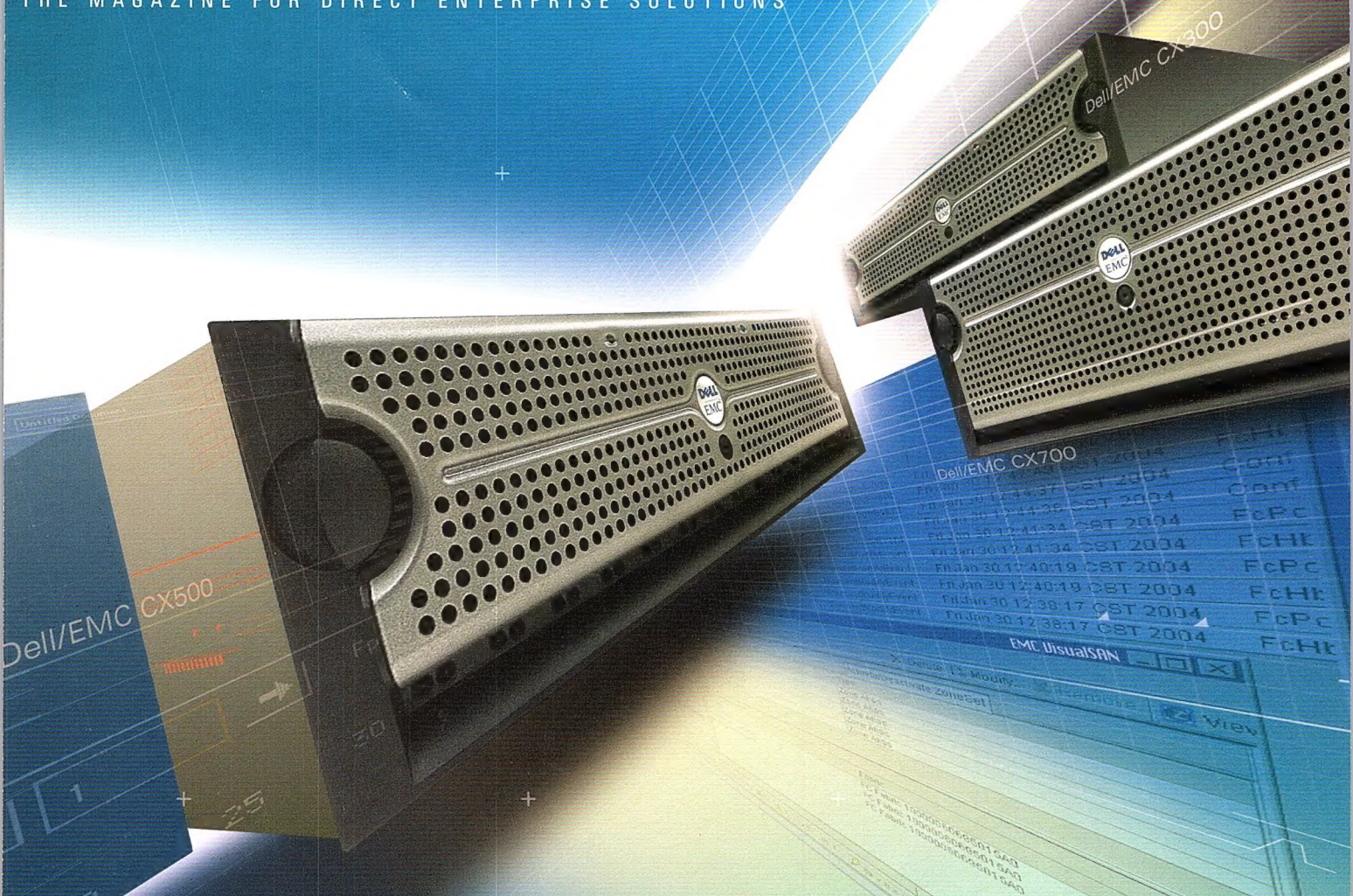


DELL™

MARCH 2004 • \$12.95

POWER SOLUTIONS

THE MAGAZINE FOR DIRECT ENTERPRISE SOLUTIONS



Managing Storage Across the Enterprise

Inside This Issue:

- Deploying modular storage with next-generation Dell/EMC arrays
- Clustering database servers in the scalable enterprise
- Exploring Intel 64-bit extension technology



UNSCRAMBLE THE PUZZLE ABOUT UNIX MIGRATION.



COSTS WILL PLUMMET AND ROI WILL SOAR.

Choose Dell coupled with Microsoft® Windows Server™ 2003 and you've found the trick to UNIX migration. It's called flexibility. And it's a combination that can give you incredible value through reduced IT costs. Plus, you'll have the agility to respond to new trends. Without question, the teaming of Dell and Microsoft can be a boon to productivity and a bear on your TCO. How's that for a better way? Find out more. Call 1-866-871-9881 or visit www.DELL.com/MSmigration and get a free business case analysis on migrating to a Dell/MS Windows Server 2003 solution.

visit www.DELL.com/MSmigration or call 1-866-871-9881
for your free UNIX migration business case analysis



Dell and the Dell logo are trademarks of Dell Inc. or its subsidiaries in the United States and other countries. Microsoft and Windows Server are registered trademarks or trademarks of Microsoft Corporation. ©2004 Dell Inc. All rights reserved. Use of the Rubik's Cube® is by permission of Seven Towns Ltd.

EDITOR'S COMMENTS

6 Mastering the Storage Domain

By Tom Kolnowski

EXECUTIVE INSIGHTS

8 Simplifying Enterprise Storage

An interview with Darren Thomas, vice president and general manager of enterprise storage systems at Dell Inc.

STORAGE ENVIRONMENT

14 Volume Shadow Copy Service Helps Build an Integrated Backup System

By Ananda Sankaran, Kevin Guinn, and Dat Nguyen

20 Managing Data Protection with Red Hat Linux and Dell PowerVault Tape Autoloaders and Libraries

By Tesfamariam Michael and Richard Goodwin

28 Protecting Business-Critical Data at Remote Offices

By Sheri Atwood and Michael Parker

31 Streamlining Backup and Recovery Operations Using Disk-based Protection

By Scott Kosciuk, Michael Parker, and Mark Thomason

34 Leveraging the Microsoft Virtual Disk Service Using QLogic SANsurfer VDS Manager

By Tim Lustig and Keith Hageman

36 Simplifying Enterprise Backup and Restore Operations Using BakBone NetVault

By Joe Gallo

SCALABLE ENTERPRISE

38 Migrating to Industry-Standard 64-bit Architectures

By John Fruehe

42 Scaling Out Microsoft Exchange 2000 Server with Dell PowerEdge 6650 Servers

By Fatima Hussain and Scott Stanford

49 Scaling Out Web Server Performance on Dell PowerEdge 6650 Servers

By David J. Morse

52 Scalable Enterprise Computing: Testing a Clustered Database on the Dell PowerEdge 6650

By Dave Jaffe, Ph.D., and Todd Muirhead

56 Optimizing Disaster Recovery Using Oracle Data Guard on Dell PowerEdge Servers

By Paul Rad, Zafar Mahmood, Ibrahim Fashho, Raymond Dutcher, Lawrence To, and Ashish Ray

59 Introducing VMware ESX Server, VirtualCenter, and VMotion on Dell PowerEdge Servers

By Dave Jaffe, Ph.D.; Todd Muirhead; and Felipe Payet

BONUS PULL-OUT POSTER

64a Dell Enterprise Storage Environment

Schematic drawings and specifications portray enterprise scenarios configured with Dell/EMC arrays, Dell PowerVault systems, and other key storage components. Visit http://www.dell.com/magazines_extras to download additional copies.

COVER STORY | PAGE 10

Managing Storage Across the Enterprise

By Sonya R. Sexton and Vicki Van Ausdall

Leveraging the flexibility of modular, next-generation Dell/EMC Fibre Channel storage arrays and Dell PowerVault network attached storage (NAS) systems, administrators can use sophisticated software to manage enterprise storage—wherever it may be.



Dell/EMC CX500 storage array

SYSTEMS MANAGEMENT

- 65** Deploying Dell OpenManage on VMware ESX Server
 By Balasubramanian Chandrasekaran, Tim Abels, Robert Wilson, and Paul Rad
- 68** Implementing Avocent AMX KVM Switches in the Dell Enterprise Solutions Engineering Group Lab
 By Mike Kosacek and Avocent Corporation
- 72** Integrating Nagios and Ganglia with Dell OpenManage Server Administrator in Linux-based Environments
 By Dan Beres, Roger Goff, and Terry Schroeder
- 76** Remotely Managing UNIX and Linux Servers Using the Dell RAC Serial/Telnet Console
 By Aurelian Dumitru
- 80** Troubleshooting Servers with Dell Remote Access Controllers
 By Jon McGary
- 83** Simplifying Enterprise Deployment of Dell Remote Access Controllers
 By Zain Kazim, Bala Beddhanan, and Alan Daughetee

- 88** System Recovery Using Windows Server 2003 on Dell PowerEdge Servers
 By Ranjith Purush, Neftali Reyes, and Edward Yardumian

LINUX ENVIRONMENT

- 94** Simplifying Linux Management with Dynamic Kernel Module Support
 By Gary Lerhaupt and Matt Domsch
- 99** Configuring and Managing Software RAID with Red Hat Enterprise Linux 3
 By John Hull and Steve Boley

NETWORK AND COMMUNICATIONS

- 103** Introduction to TCP Offload Engines
 By Sandhya Senapathi and Rich Hernandez
- 108** Improving Quality of Service Using Dell PowerConnect 6024/6024F Switches
 By Marvell Semiconductor
- 113** 10 Gigabit Ethernet Helps Relieve Network Bottlenecks for Bandwidth-Intensive Applications
 By Matt W. Baker and Wu-chun Feng

ADVERTISER INDEX

Avocent Corporation	19	Microsoft Corporation	4-5
Dell Inc.	C2, 15, 23, 39, 43, 47	Novell, Inc.	C4
Emulex Corporation	3	Oracle Corporation	C3
McDATA Corporation	27	VERITAS Software Corporation	7

EDITORIAL

EDITOR-IN-CHIEF | Tom Kolnowski

MANAGING EDITOR | Debra McDonald

SENIOR EDITORS | Liza Graffeo, Lori Kennedy, Cathy Luo, Vicki Van Ausdall

CONTRIBUTING AUTHORS | Tim Abels, Sheri Atwood, Matt W. Baker, Bala Beddhanan, Dan Beres, Steve Boley, Balasubramanian Chandrasekaran, Chris Croteau, Alan Daughetee, Matt Domsch, Aurelian Dumitru, Raymond Dutcher, Ibrahim Fashho, Wu-chun Feng, John Fruehe, Joe Gallo, Roger Goff, Richard Goodwin, Michael Grant, Kevin Guinn, Keith Hageman, Rich Hernandez, John Hull, Fatima Hussain, Dave Jaffe, Ph.D., Zain Kazim, Mike Kosacek, Scott Kosciuk, Gary Lerhaupt, Paul Luse, Tim Lustig, Zafar Mahmood, Jon McGary, Tesfamarium Michael, David J. Morse, Todd Muirhead, Dat Nguyen, Michael Parker, Felipe Payet, Ranjith Punshi, Paul Rad, Ashish Ray, Neftali Reyes, Ananda Sankaran, Terry Schroeder, Sandhya Senapathi, Sonya R. Sexton, Scott Stanford, Mark Thomason, Lawrence To, Vicki Van Ausdall, Robert Wilson, Edward Yardumian

ART

ART DIRECTOR | Mark Mastroianni

DESIGNERS | Phu Tran, Cynthia Webb

ILLUSTRATOR | Cynthia Webb

COVER DESIGN | Phu Tran

CONTRIBUTING PHOTOGRAPHER | Lee Kirgan

ONLINE

WEB PRODUCTION | Brad Klenzendorf

SUBSCRIPTION SERVICES

Subscriptions are free to qualified readers who complete the online subscription form or the subscription reply card found in each issue. To sign up as a new subscriber, renew an existing subscription, change your address, or cancel your subscription, submit the online subscription form at www.dell.com/powersolutions_subscribe, return the subscription reply card by surface mail, or fax the subscription reply card to +1 512.283.0363. For subscription services, please e-mail us_power_solutions@dell.com.

ABOUT DELL

Dell Inc., headquartered in Round Rock, Texas, near Austin, is the world's leading direct computer systems company. Dell is one of the fastest growing among all major computer systems companies worldwide, with approximately 40,000 employees around the globe. Dell uses the direct business model to sell its high-performance computer systems, workstations, and storage products to all types of enterprises. For more information, please visit our Web site at www.dell.com.

Dell, Latitude, OpenManage, OptiPlex, PowerConnect, PowerEdge, PowerVault, Precision—Dell Inc.; Adaptec—Adaptec, Inc.; Altiris—Altiris, Inc.; ANSI—American National Standards Institute, Inc.; OSCAR, Outlook—Apex PC Solutions, Inc.; AMWorks, AMX—Avocent—Avocent Corporation; APM—Application Plugin Module; BakBone, NetVault, SmartClient—BakBone Software, Inc.; Broadcom—Broadcom Corporation; Brocade—Brocade Communications Systems, Inc.; Cisco—Cisco Systems, Inc.; Access Logic, EMC, EMC Proven, MirrorView, Navisphere, PowerPath, SAN Copy, SnapView, Symmetrix, VisualSAN, Visual SRM—EMC Corporation; BIG-IP—F5 Networks, Inc.; Force10—Force10 Networks, Inc.; HyperTerminal—Hilgraeve, Inc.; IEEE—Institute of Electrical and Electronics Engineers, Inc.; Intel, Itanium, Pentium, Xeon—Intel Corporation; AIX, Chipkill, DB2, eServer, IBM, Informix, ServeRAID, xSeries—International Business Machines Corporation; Linux—Linux Torvalds; LSI Logic—LSI Logic Corporation; Active Directory, Microsoft, MS-DOS, Outlook, SQL Server, Windows, Windows NT, Windows Server—Microsoft Corporation; MySQL—MySQL AB; Nortel Networks—Nortel Networks; NetWare, Novell—Novell, Inc.; Oracle, Oracle9i—Oracle Corporation; PCI, PCI Express, PCI-X—PCI SIG Corporation; OLogic, SANsurfer—OLogic Corporation; Red Hat, Red Hat Certified Engineer, RHCE, RPM—Red Hat Software, Inc.; SAP—SAP Aktiengesellschaft; SPEC, SPECint, SPECweb—Standard Performance Evaluation Corporation; Java—Sun Microsystems, Inc.; Sybase—Sybase, Inc.; UNIX—The Open Group; VERITAS Backup Exec, ExecView, VERITAS NetBackup, VERITAS Storage Replicator, VERITAS—VERITAS Software Corporation; ESX Server, VMotion, VMware—VMware, Inc.; TapeWare, Yosemite Technologies—Yosemite Technologies, Inc.; Zeus Web Server—Zeus Technology Ltd. Other company, product, and service names may be trademarks or service marks of others.

Dell Power Solutions is published quarterly by the Dell Product Group, Dell Inc. *Dell Power Solutions*, Mailstop 8456, Dell Inc., One Dell Way, Round Rock, TX 78682, U.S.A. This publication is also available online at www.dell.com/powersolutions. No part of this publication may be reprinted or otherwise reproduced without permission from the Editor-in-Chief. Dell does not provide any warranty as to the accuracy of any information provided through *Dell Power Solutions*. Opinions expressed in this magazine may not be those of Dell. The information in this publication is subject to change without notice. Any reliance by the end user on the information contained herein is at the end user's risk. Dell will not be liable for information in any way, including but not limited to its accuracy or completeness. Dell does not accept responsibility for the advertising content of the magazine or for any claims, actions, or losses arising therefrom. Goods, services, and/or advertisements within this publication other than those of Dell are not endorsed by or in any way connected with Dell Inc.

ON THE FRONT COVER | The latest Dell/EMC CX300, CX500, and CX700 Fibre Channel storage arrays, flanked by EMC VisualSAN management software. Design by Phu Tran. Photos by Lee Kirgan.

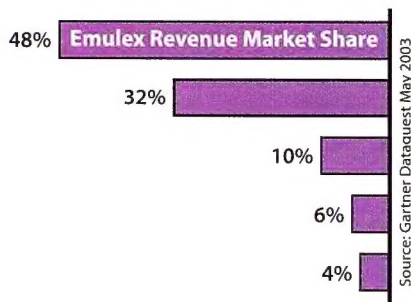
Copyright © 2004 Dell Inc. All rights reserved. Printed in the U.S.A.
 March 2004



Emulex.

We network storage

Emulex HBAs. Proven choice of the world's leading server and storage providers.



Emulex Fibre Channel host bus adapters are well known for their high performance, reliability and robust interoperability. Additionally, unique capabilities such as operating system driver compatibility across the entire product line and firmware upgradeability allow Emulex to deliver on the promise of storage area networks by simplifying SAN management and protecting customer investments.*

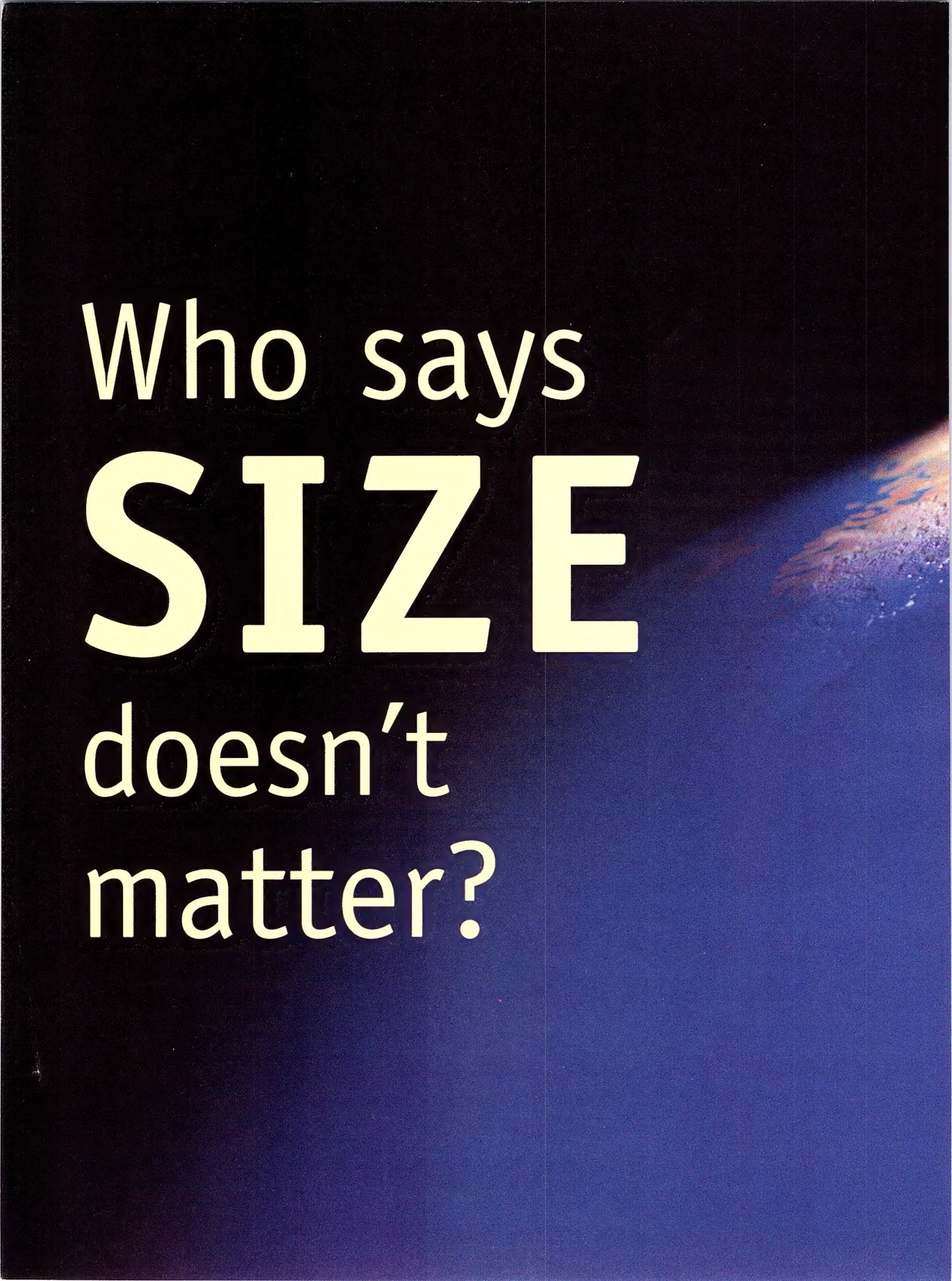
Emulex, #1 in Fibre Channel HBAs.



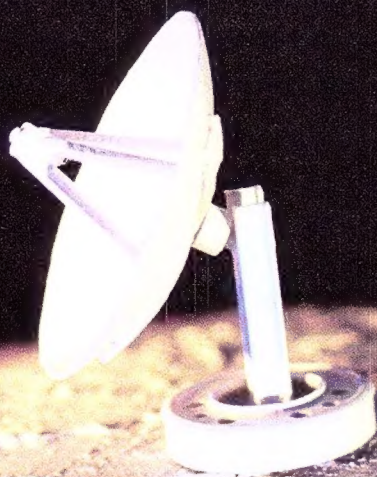
EMULEX™

We network storage

*See IDC white paper at www.emulex.com
04-053



Who says
SIZE
doesn't
matter?



When your IT organization has
terabytes of data to process,

**YOU'VE
GOT TO
THINK BIG.**

Microsoft and Dell understand this.

A proven combination, Microsoft®
SQL Server running on Dell™ PowerEdge™
servers can handle billions of records with
impressive performance. More companies
than ever entrust their biggest jobs to this
powerful enterprise solution—making SQL
Server on Dell hardware a fast-growing
solution for very large databases.

Of course bigger isn't always better. That's
why this major processing power is available
for a minimal total cost of ownership.

To learn more, visit www.microsoft.com/sql
and www.dell.com/sql.



Microsoft
SQL Server 2000
Enterprise Edition



© 2004 Microsoft Corporation and Dell Inc. All rights reserved. Microsoft is a registered trademark of Microsoft Corporation. Dell, the Dell logo, and PowerEdge are trademarks of Dell Inc. Other company, product, and service names may be trademarks or service marks of others.



Mastering the Storage Domain

As storage requirements escalate across the board, decisions about how to deploy, scale, and—most importantly—*manage* storage resources can assume the weight of strategic business initiatives. To master the enterprise storage domain, administrators must think out of the box, beyond the individual storage systems in geographically dispersed data centers. Taking a modular approach can enable IT organizations to increase capacity cost-effectively by capitalizing on powerful networked storage arrays and smart storage management software.

The first 2004 issue of *Dell Power Solutions* delves into the modular approach to storage management, capacity planning, and performance tuning. Among the highlights is our cover story, “Managing Storage Across the Enterprise.” Featured coverage includes:

- **Optimized storage resource utilization:** Learn how to meet skyrocketing capacity requirements within tight budget constraints by deploying the appropriate software modules in the storage environment.
- **Improved data availability and disaster recovery:** Use smart software and point-and-click graphical user interfaces (GUIs) to produce multiple copies of critical data anytime, anywhere—locally or remotely.
- **New Dell™/EMC® CX300, CX500, and CX700 storage arrays:** Explore the new modular storage arrays, which combine the versatility of industry-standard building blocks with high-level performance, scalability, and manageability.
- **Storage environment pull-out poster:** Scope out the storage landscape of Dell/EMC arrays, Dell PowerVault™ systems, and other key storage elements in scenarios such as corporate headquarters, central backup, and disaster recovery. Plus: a large schematic diagram of the new Dell/EMC CX500 storage array enclosure flanked by related storage components.

In addition, this issue presents many informative articles in the Storage Environment section, including how to leverage

the Volume Shadow Copy Service in Microsoft® Windows Server™ 2003 and how to manage Dell PowerVault tape technology in Linux® environments. Articles in the Scalable Enterprise section explore the latest release of VMware® ESX Server™ software, industry-standard clustered database servers, and Microsoft Exchange scalability. On the topic of networking, learn more about quality of service (QoS) mechanisms in the latest generation of Dell PowerConnect™ 6024/6024F Ethernet switches.

Dell Power Solutions has always offered both print and online versions, but beginning with this issue we will be publishing online-exclusive content in a new Magazines Extras section on our Web site—bookmark http://www.dell.com/magazines_extras for direct access. From there, you can browse value-added content from selected articles and download additional copies of the storage environment pull-out poster—including both standard 20.75 × 30.75-inch size and ISO A1 form factor.

Last of all, you will notice a new signature at the end of this editorial. It is with great expectation and excitement that I take the reins of *Dell Power Solutions* and our sister publication, *Dell Insight*. After more than six years at Dell in a variety of capacities, I feel highly privileged to be settling into this chair. Please help keep us focused on what is most important to your IT needs by continuing to send your comments our way.

Happy reading!

Tom Kolnowski
Editor-in-Chief
tom_kolnowski@dell.com
www.dell.com/powersolutions

VERITAS™

V I S I O N

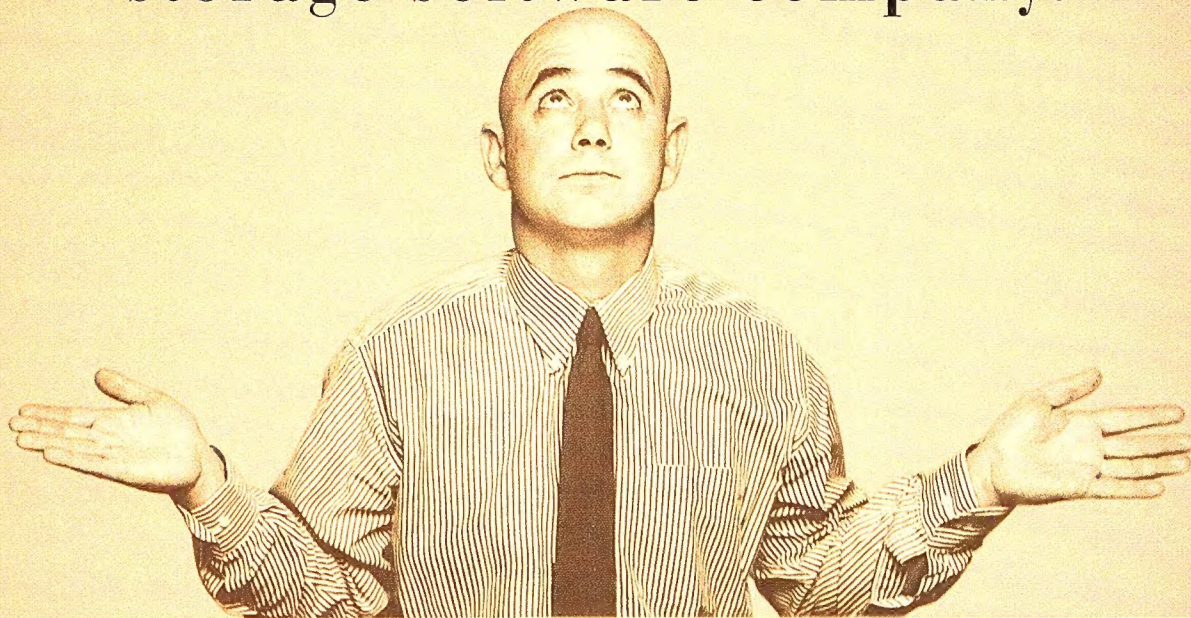
MAY 3 - 7, 2004

VENETIAN HOTEL | LAS VEGAS, NV
www.veritas.com/vision/register.html

UTILITY. NOW.

The
“NETBACKUP 5:
TAPE, DISK,
WHATEVER”

storage software company.



NetBackup, the fastest tape backup software, is now the fastest disk backup software. For on-line, near-line or far-line trust NetBackup 5. veritas.com

VERITAS™

Copyright © 2004 VERITAS Software Corporation. All rights reserved. VERITAS and the VERITAS Logo are trademarks of VERITAS Software Corporation Reg. U.S. Pat. & Tm. Off.

Simplifying Enterprise Storage



Darren Thomas, vice president and general manager of enterprise storage systems at Dell Inc., discusses the Dell™ strategy for building a modular, industry-standard storage infrastructure—including the new Dell/EMC® CX series arrays, which can help organizations of all sizes consolidate storage resources and shorten recovery times.

Managing storage has become a pressing concern for IT organizations facing exponential data growth with limited budgets and staff. As enterprise requirements for scalable capacity and 24/7 service intensify, IT managers are exploring ways to provide high availability for business-critical applications and to protect valuable data assets throughout the organization.

Dell is committed to helping enterprises implement modular, scalable storage environments that allow administrators to add incremental capacity quickly and flexibly to meet changing business needs. The Dell™ approach leverages industry-standard building blocks and integrated storage management software that enables organizations of all sizes to create cost-effective networked storage environments. Darren Thomas, vice president and general manager of enterprise storage systems at Dell Inc., explains how the Dell storage strategy resonates with customers—as evidenced by the recent Dell ascension to the No. 4 position in the total disk storage systems market.¹

What does the Dell storage product portfolio look like today?

More organizations than ever are choosing Dell as their storage supplier. We view this as a positive response to our broad product offerings—ranging from our Dell PowerVault™ SCSI enclosures and PowerVault tape systems to our Dell/EMC® fabric attached storage products and award-winning² PowerVault network attached storage (NAS).

How does the Dell and EMC partnership affect product offerings?

The foundation of the Dell/EMC partnership is the shared strength of the Dell direct model—which delivers built-to-order, standards-based hardware components at a low cost—and EMC leadership in the area of storage management software. By combining the core competencies of each company, we can bring better products to market faster, drive customer-focused product development, and avoid duplication of engineering and marketing efforts.

To what do you attribute recent Dell growth in the SAN storage market?

Networked storage systems are critical to the scalable enterprise. Last year Dell experienced considerable growth in storage area network (SAN) deployments across a broad range of markets. We believe this success to be a result of Dell and EMC efforts to make enterprise-class SAN storage more accessible to all tiers of the market—providing networked storage for organizations that previously could not afford it.

In the area of services, Dell offers enterprise customers robust, customized Gold-level Dell/EMC Premier Enterprise Support Services—including a single point of contact for 24/7 support, remote storage system monitoring, and more. Storage planning, deployment, and training services also are available—including one of the first skills-based storage networking professional certification programs—to give administrators hands-on experience in building and managing SANs.

¹ IDC's *Worldwide Disk Storage Systems Quarterly Tracker* for Q303, December 4, 2003.

² For example, Dell PowerVault 770N and 775N NAS servers received a Reader's Choice Award from *Windows & .NET Magazine* (see "Network Attached Storage" in *Windows & .NET Magazine*, September 15, 2003, <http://www.winnetmag.com/articles/print.cfm?articleID=40105>). In addition, the Dell PowerVault 775N system received an *InfoWorld* 2003 Readers' Choice Award ("2003 Readers' Choice Awards" by Leslie T. O'Neill in *InfoWorld*, July 28, 2003, http://www.infoworld.com/pdf/special_report/ReaderChoice.pdf) and a *Network Computing* Editor's Choice Award ("First-Class NAS" by Steven Schuchart Jr. in *Network Computing*, August 21, 2003, <http://www.networkcomputing.com/1416/1416f32.html>).

What strengths do the new Dell/EMC storage arrays offer?

The next-generation Dell/EMC CX300, CX500, and CX700 storage arrays provide significantly more power and flexibility than the previous CX series family at a comparable cost. Improving on the previous generation—including the highly rated³ CX200 array—the new CX series arrays enable organizations to consolidate storage onto fewer arrays. This helps to lower capital and operating expenses and simplify what is arguably the most expensive aspect of storage ownership—management.

CX300, CX500, and CX700 arrays are fully supported by the EMC Navisphere® Management Suite, which offers backward compatibility with previous Dell/EMC arrays. Navisphere—as well as optional tools such as EMC VisualSAN®—simplify storage administration by presenting a unified view of enterprise storage resources from a single console. Centralized management lets organizations efficiently share global disk resources among arrays, and sophisticated data-copy applications in the suite—such as EMC SnapView™ and SAN Copy™—reduce manual management tasks and facilitate disaster recovery planning by automating the production of data copies.

The modular design of the CX series arrays enables organizations to scale capacity incrementally and nondisruptively as enterprise storage needs change. Using Navisphere, administrators can provision additional capacity online while arrays remain fully functional—helping to keep business-critical networked applications highly available. The CX series arrays can be deployed in SAN, NAS, and direct attach storage (DAS) configurations so that organizations can share enterprise storage flexibly across different types of connections.

How does Dell support emerging serial disk protocols?

The transition of disk protocols from parallel to serial architecture in the enterprise is well underway—largely in response to the increased requirements of today's bandwidth-intensive applications. Originally conceived as a desktop architecture, Serial ATA (SATA) offers major performance advantages over parallel ATA and provides features such as hot-plug drive swapping that have led IT organizations to consider SATA for enterprise uses. Another serial technology—Serial Attached SCSI (SAS)—can help reduce overhead for enterprise systems and lower bus bandwidth. SAS is expected to be highly scalable and offers backward compatibility with legacy SCSI drivers and software. As these technologies mature, Dell will consider products that help enterprises take advantage of the performance gains and cost-effectiveness promised by serial technology.

What role is iSCSI expected to play in the storage industry?

Dell is excited about the potential for Internet SCSI (iSCSI) to lower the cost of network connections, making networked storage more widely available. By using IP networks to link data storage devices and transfer data, iSCSI enables administrators to deploy SANs in local area network (LAN), wide area network (WAN), and metropolitan area network (MAN) configurations. Dell will continue to consider iSCSI as well as other low-cost storage technologies to meet customer needs.


What roles do NAS and tape play in today's enterprise storage?

Both NAS and tape environments will continue to be important in the enterprise storage domain. While Dell offers stand-alone NAS enclosures, NAS and file services will become increasingly available through gateways connecting to the SAN. In addition, Dell sees NAS products based on Microsoft® Windows® Storage Server and other operating systems gaining ground, not only in lower-priced entry-level configurations but also in many high-performance enterprise computing environments.

Tape continues to be a strong business for Dell; despite the retrieval performance benefits of disk technology, disk is not likely to replace tape. Tape is evolving from a secondary to a tertiary role in enterprise storage and will continue to be a key part of enterprise data storage infrastructures because of its long life, durability, and portable nature—not only for long-term and off-site storage but also for disaster recovery and business continuance planning. Dell also recommends that customers continue to consider Dell/EMC DAE2 enclosures with the ATA disk option as an effective alternative to tape for many data archival needs.

What does the future of storage hold for Dell?

At Dell, we believe that strong interoperability and integration are essential to the future of affordable storage. To this end, Dell is focused on driving storage standards that will help administrators transfer data from disk to disk more efficiently and easily share resources among heterogeneous storage systems—allowing organizations to take advantage of scalable storage building blocks. Dell is working closely with the Storage Network Industry Association (SNIA) to help achieve a standard data format. Both Dell and EMC fully endorse SNIA SMI-S (Storage Management Initiative Specification), the emerging industry standard for storage management.

Dell believes that its efforts to meet enterprise storage needs cost-effectively and to support interoperability through standardization will continue to have broad appeal for IT organizations—as the increased Dell presence in the storage market indicates. 

³ The Dell/EMC CX200 storage array received the highest rating of "Excellent" and scored a perfect 10 for management ("Entry-Level SAN Arrays Square Off" by Paul Venezia in *InfoWorld*, July 7, 2003). For more information, see http://www.dell.com/downloads/global/products/pvaul/en/pvaul_cx200_award.pdf.

Managing Storage Across the Enterprise

Leveraging the flexibility of modular, next-generation Dell™/EMC® Fibre Channel storage arrays and Dell PowerVault™ network attached storage (NAS) systems, administrators can use sophisticated software to manage storage efficiently throughout the enterprise. This article explores how powerful software management features can enable administrators to respond quickly and cost-effectively to business demands—helping to ensure the availability of mission-critical data while reducing the complexity of enterprise storage management.

BY SONYA R. SEXTON AND VICKI VAN AUSDALL

To put valuable data assets to best use in today's aggressive business climate, IT organizations must find efficient, cost-effective ways to scale storage capacity and performance—and accommodate increasingly diverse data availability requirements. Modular Dell™/EMC® Fibre Channel storage arrays and Dell PowerVault™ network attached storage (NAS) systems are based on industry-standard components that enable administrators to address immediate business concerns flexibly and grow incrementally as needed.

Largely because of advanced software that provides new functionality in areas of capacity provisioning, storage resource management (SRM), and storage area network (SAN) infrastructure management, modular storage components can now fulfill enterprise-class storage requirements. The latest Dell/EMC CX series arrays incorporate several high-end capabilities that help IT organizations administer geographically dispersed storage farms

efficiently throughout the enterprise (see “Modular building blocks: Dell/EMC CX300, CX500, and CX700 storage arrays” for more information). In addition, Microsoft® Windows® Storage Server 2003–based Dell PowerVault NAS systems have become sophisticated enough to be used in high-end enterprise NAS deployments. These modular components create the basis for an integrated approach to storage management that can help simplify system administration, optimize resource utilization, and improve data availability and disaster recovery.

Simplifying system administration

As enterprises grow, storage infrastructures rapidly become more complex. This complexity can overburden existing administrative resources and drastically increase total cost of ownership (TCO). Centralized storage resources can improve management efficiency considerably, and SANs are an effective venue for consolidating

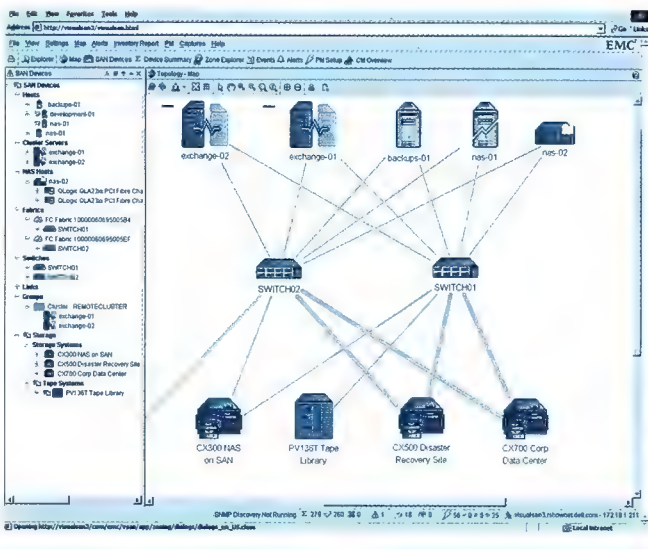


Figure 1. Remote storage management using EMC VisualSAN

control. EMC VisualSAN® software allows administrators to manage growth and change with less effort by providing high-level visibility into a Dell/EMC SAN infrastructure (see Figure 1). VisualSAN monitors the health of hardware devices on the SAN, helping administrators identify problem conditions before they cause costly, unplanned downtime. Advanced performance management features—including real-time and historical performance reports—further reduce the risk of downtime by helping administrators proactively tune performance and optimize throughput (see Figure 2).

By providing a centralized, graphical view of the storage infrastructure and powerful tools to manage SAN devices, VisualSAN can help free IT staff from contending with repetitive management tasks and multiple incompatible software programs. The tool also helps simplify configuration changes and expedite the deployment of new equipment, all of which can improve productivity and streamline storage management in expanding enterprise environments. VisualSAN supports PowerVault NAS systems as well as storage arrays. Increasingly, IT departments are deploying NAS systems within SAN environments, which enables NAS to become an integrated component within the total enterprise storage infrastructure. As organizations begin to provide NAS and file services through gateways connected to the SAN, stand-alone NAS systems may become less common.

Optimizing resource utilization

One of the greatest challenges facing organizations tasked with managing enterprise storage is meeting escalating capacity requirements within tight budget constraints. To address such considerations, all CX series Dell/EMC storage arrays include the EMC Navisphere® Management Suite—a comprehensive set of storage management tools that give administrators central control of Dell/EMC arrays.

By retaining a consistent look-and-feel through many versions, Navisphere helps enterprises capitalize on existing IT expertise and avoid potentially expensive and disruptive retraining. Navisphere also helps protect existing hardware investments by providing backward compatibility with legacy Dell/EMC arrays. From the Navisphere console, administrators can provision additional capacity online, both at the primary site and at remote locations. Online provisioning provides administrators with the flexibility to increase capacity incrementally, in response to specific business needs, instead of overprovisioning simply to ensure that unpredictable organizational capacity demands can be met. For example, using Navisphere, administrators can allocate capacity to new or existing logical storage units (LUNs) without interrupting network applications that must access disk resources in the array. In this way, online provisioning helps ensure application availability and minimize planned and unplanned downtime.

Furthermore, the most recent version of Navisphere enables administrators to create consolidated groups of LUNs, or *metaLUNs*, which help improve performance by allowing a volume to span several drives. More importantly, from a cost perspective metaLUNs can increase the utilization of existing storage resources by allowing administrators to provision additional capacity from available disk resources anywhere in the array.

Navisphere also enables administrators to configure hot spares and copy repositories that can be globally shared throughout the array. This approach avoids the requirement to set aside discrete disk resources for individual disk groups, and enables organizations to meet storage requirements using fewer physical disk drives. Using shared global disk resources for copy space, with unique tools for right-sizing repositories, can make implementing copy operations easier and significantly more affordable.

To enhance Navisphere functionality, EMC VisualSRM™ storage resource management software provides a view of the capacity

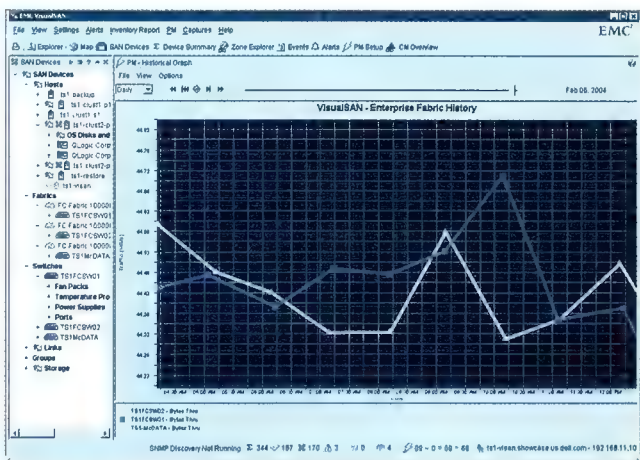


Figure 2. VisualSAN Performance Monitor replay feature

MODULAR BUILDING BLOCKS: DELL/EMC CX300, CX500, AND CX700 STORAGE ARRAYS

The newest family of Dell/EMC modular storage arrays combines the versatility of industry-standard building blocks with performance, scalability, and manageability at a price comparable to that of previous-generation arrays. Administrators can implement the Dell/EMC CX300, CX500, and CX700 arrays in storage area network (SAN), direct attach storage (DAS), or network attached storage (NAS) configurations, in any combination. This flexibility allows IT organizations to share storage across several types of connections. Designed for demanding workloads, the new CX series arrays allow administrators to consolidate storage onto fewer high-performance modules, which can simplify systems management and reduce capital and operating costs.

Modular Dell/EMC CX300, CX500, and CX700 storage arrays enable organizations to add capacity as needed through nondisruptive hardware and software upgrades that help keep data highly available for 24/7 operations (see "Optimizing resource utilization"). The new Dell/EMC CX series arrays help extend the life of hardware investments by creating a storage infrastructure that can grow incrementally as business requirements evolve. The Dell/EMC CX300, CX500, and CX700 arrays are fully compatible with existing CX series components such as disks, disk array enclosures (DAEs), and software.

The EMC Navisphere Management Suite—an integrated set of software tools that enables centralized storage management of Dell/EMC arrays as well as convenient Web-based access—is backward compatible across legacy CX series arrays. The highly integrated nature of the data-copy applications in the Navisphere suite, as well as integration with popular third-party database and messaging applications, helps streamline

The new Dell/EMC CX series arrays help extend
the life of hardware investments by creating
a storage infrastructure that can grow incrementally
as business requirements evolve.



Dell/EMC CX700
storage array



Dell/EMC CX500 storage array

common tasks, reduce setup time for data replication jobs, and automate disk-based backup. The powerful management and performance-tuning capabilities enabled by Navisphere and other storage management software, such as EMC VisualSAN and EMC VisualSRM, help improve administrator productivity so that existing IT staff can manage increasing enterprise storage capacity easily.

The new Dell/EMC arrays are offered in three sizes to address a variety of enterprise needs. The midrange Dell/EMC CX500 supports up to 120 disk drives, 4 GB of cache memory, and four front-end host ports—making it suitable for departmental server and Tier 2 purposes as well as a broad range of uses in distributed organizations:

- High-performance enterprise applications
- High-speed content delivery such as video streaming
- Mirrored disaster recovery
- Online transaction processing (OLTP) and Web serving
- Data warehousing
- Graphics-intensive applications
- Messaging applications
- Customer relationship management (CRM) applications

The CX700 is a high-performance, mid-tier platform that provides the same functions as the CX500, but is designed with more powerful processing resources for heavy database and OLTP applications. The CX700 features eight back-end Fibre Channel buses to support up to 240 drives, 8 GB of cache memory, and eight front-end host ports.

Also designed for business-critical and production environments that require leading-edge performance and scalability, the entry-level CX300 supports up to 60 disk drives, 2 GB of cache memory, and four front-end host ports.

utilization of hosts, file systems, and storage devices on the SAN, providing administrators with an integrated outlook on how resources are being used. VisualSRM displays key information such as which volumes have the highest usage levels. The advanced VisualSRM reporting feature includes application programming interface (API)-level integration with leading database and messaging applications such as those from Oracle and Microsoft. The information from these reports can help administrators better understand growth patterns and data placements to plan effectively for future capacity.

Improving data availability and disaster recovery

EMC provides centralized, well-integrated control of data-copy service applications through the Navisphere graphical user interface (GUI). In enterprise storage environments, the ability to produce multiple copies of data—both internally within an array and externally at a remote data site—allows administrators to increase data availability and enable disaster recovery. EMC offers a comprehensive range of data-copy services through applications such as EMC SnapView™, which creates point-in-time snapshots and clones of original data; EMC MirrorView™, which provides synchronous mirroring of critical data between Dell/EMC systems; and EMC SAN Copy™, which copies entire volumes over a high-speed SAN infrastructure or wide area network (WAN) to Dell/EMC and heterogeneous storage systems.

Of course, data-copy services are nothing new in enterprise storage environments. Remote mirroring has become fundamental to many disaster recovery plans, and point-in-time snapshots are a common way to back up data online without creating overhead on production database servers. However, EMC copy applications enable advanced capabilities through API-level integration with Navisphere, each other, and leading database and messaging applications. These copy applications can automate the creation of snapshots, which can be used for offline processing, and clones, which can be used as business continuance volumes.

Advanced versioning capabilities available with EMC copy applications can provide the benefits of secondary processing, data mobility, and disaster recovery. By allowing copy relationships to be set up for several LUNs in each array, administrators can secure more data assets, glean greater business value and competitive advantage from existing assets, and respond to growth more flexibly. In addition, advanced versioning enables multiple copy relationships to be defined for each LUN, allowing administrators to fulfill diverse requirements for point-in-time and mirror images.


EMC copy applications also include powerful tools for fast, automatic resynchronization of both mirrored volumes and point-in-time images following planned or unplanned copy session interrupts. To control the priority of resynchronization and the impact of resynchronization on production operations, EMC copy applications integrate a throttling mechanism. Also, a persistence

Planning tools: Visit the new Dell Interactive Storage Explorer at www.dell.com/storageexplorer for a guided tour of modular storage systems and software that can help organizations respond flexibly to changing business needs. Also, don't miss the companion poster to this article, which maps out key storage components for enterprise headquarters, central backup, and disaster recovery deployments (see page 64a).

feature is available to preserve consistent point-in-time images following power failures.

Supporting industry standards for interoperability

To meet diverse and often unpredictable storage needs, the modular approach helps IT organizations add capacity incrementally and manage data cost-effectively. Dell and EMC are helping to advance interoperability efforts by endorsing the adoption of standards in the storage industry. Dell is currently chairing a committee within the Storage Network Industry Association (SNIA) to develop a standard data format that will help administrators transfer information easily among heterogeneous systems. At the same time, EMC fully supports SNIA SMI-S (Storage Management Initiative Specification), the emerging industry standard for storage management.

By making it easier for storage hardware and management software from different vendors to function smoothly together, storage management standards promise to simplify system administration and improve data availability. By facilitating an integrated, modular approach, Dell and EMC products can help administrators deploy storage capacity quickly and economically, wherever enterprises need it most. 

Sonya R. Sexton (sonya_r_sexton@dell.com) is a storage systems competitive intelligence analyst for the Enterprise Marketing Operations Group at Dell. She has an M.F.A. from the University of Alabama.

Vicki Van Ausdall (vicki@tdagroup.com) is a senior editor at *Dell Power Solutions*, with 12 years of experience as a technical writer and editor for various high-tech publications in the San Francisco Bay Area. She has a B.A. in English Literature from Hamilton College.

FOR MORE INFORMATION

Dell/EMC CX300, CX500, CX700, and Dell PowerVault NAS:

<http://www.dell.com/storage>

EMC VisualSAN:

http://www.emc.com/products/software/visual_san/visual_san.jsp

EMC VisualSRM:

http://www.emc.com/products/software/visual_srm/visual_srm.jsp

EMC Navisphere Management Suite: <http://www.dell.com/downloads/global/products/pvaul/en/navispheremanagementsuite.pdf>

Volume Shadow Copy Service

Helps Build an Integrated Backup System

The Microsoft® Windows Server™ 2003 operating system and its enhanced storage management features—particularly the Volume Shadow Copy Service (VSS)—can provide the framework for an integrated backup system based on industry-standard components such as Dell™ PowerEdge™ servers, Dell/EMC® storage arrays, Dell PowerVault™ tape libraries, and VERITAS Backup Exec™ software. This article explores how VSS can enable these components to interoperate, helping to reduce the time and complexity of online backups.

BY ANANDA SANKARAN, KEVIN GUINN, AND DAT NGUYEN

To be competitive in the ever-changing global economy, many enterprises must provide uninterrupted access to key services such as customer databases, e-commerce applications, and corporate messaging systems. Taking systems offline to back up data is no longer a viable option—but creating reliable backup and recovery services for enterprise systems that must remain available 24/7 can be a challenge for IT administrators. This article presents an example scenario explaining how IT administrators can integrate the backup process for a high-availability Microsoft® Exchange Server 2003 messaging system using standards-based enterprise components and the Volume Shadow Copy Service (VSS), a new file-system feature included in the Microsoft Windows Server™ 2003 operating system.

When a key business service such as Microsoft Exchange Server 2003 is online, several files are either open or undergoing changes. The large volume and dynamic nature of the data make an accurate and reliable backup difficult to perform. Creating point-in-time copies of storage volumes, such as the snapshots taken by EMC® SnapView™ software, is one way administrators can reduce the time required to back up data. However, because the

underlying data is constantly changing, I/O operations to a storage volume must be paused to ensure that the data is consistent before a reliable snapshot can be taken. The challenge for system administrators is finding a way to integrate and automate the processes of pausing I/O, creating snapshots, resuming I/O, and performing backups from the snapshot data while the original data and service are restored to operation.

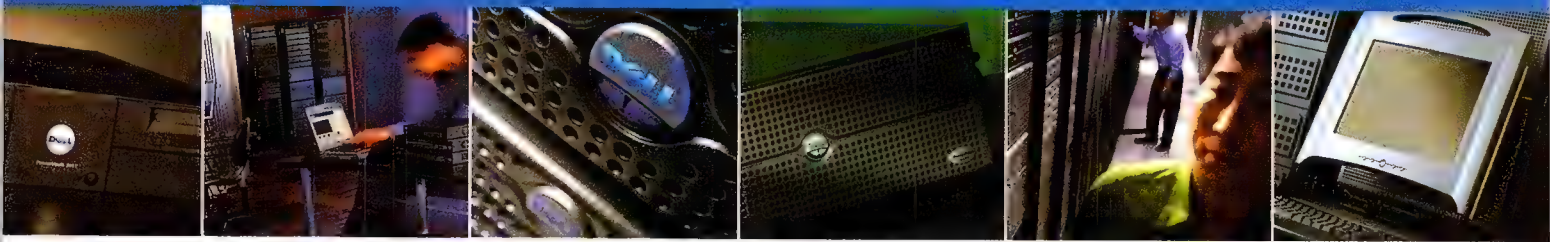
VSS provides a framework and an application programming interface (API) that enables administrators to automate several processes related to creating snapshots and performing backups. Because it allows third-party storage hardware, backup software, and application software to interoperate within the VSS framework, VSS can enable administrators to streamline the storage management process—helping to reduce the time, complexity, and cost of online backups considerably.

VSS on Windows Server 2003: Shadow copies

VSS allows system administrators to create software-driven snapshots, or *shadow copies*, of a volume that can be shared as a network resource. The shadow copy can be created on a reserved area that resides in the same volume

Dell Enterprise Webcast Series

Sponsored by *Dell Power Solutions*



Experts are coming to talk to you right at your desk. Join the Dell Enterprise Webcast series to learn about the latest technologies and current IT issues impacting enterprises.

First Webcast:

Data Archiving Best Practices for Regulatory Compliance
Tuesday, April 20, 2004 – 3:00 p.m. (CST)

Many organizations are scrambling to reevaluate their data management strategies because of a renewed focus on regulatory compliance and increasing requirements for legal discovery involving electronic data. Hear best practices from an industry expert and learn how Dell can help enterprises meet these challenges with cost-efficient solutions, designed to scale incrementally and leverage previous investments.

Please visit <http://www.dell.com/dellevents> to register.



DELL
POWER
SOLUTIONS
THE WAY TO THE FUTURE OF BUSINESS

Data Archiving Solutions. Easy as **DELL**™

Visit www.dell.com for more information.

Dell is not responsible for errors in typography or photography, or omissions. Dell and the Dell logo are trademarks of Dell Inc. Other trademarks and trade names may be used in this document to refer to either the entities claiming the marks and names or their products. Dell disclaims proprietary interest in the marks and names of others. © Copyright 2004 Dell Inc. All rights reserved. Reproduction in any manner whatsoever without the express written permission of Dell Inc. is strictly forbidden. For more information contact Dell. February 2004, Kolar

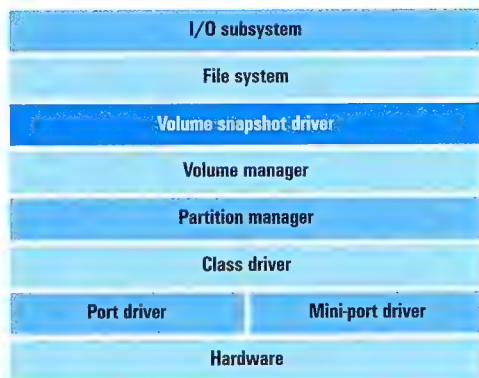


Figure 1. Windows Server 2003 layered driver model

as the data, or it can be created on a reserved area that resides in a different volume. This native shadow copy functionality uses a copy-on-first-write method, which makes a duplicate of any block that has been changed since the last shadow copy was created. Only changed file blocks are copied to the snapshot just before the change, not the entire volume.

After a snapshot is created, VSS oversees all activities on the source volume, including writes and reads:

- **Writes:** When an application or a user modifies a file or sector on the source volume, VSS identifies which blocks have been affected. If a shadow copy exists and the block has not been modified since the most recent shadow copy was created, VSS first saves the original data in the shadow copy volume and then writes the change to the source volume. If the block has been modified previously, VSS writes the change directly to the source volume.
- **Reads:** When an application or a user reads a file or sector from the source volume, VSS typically serves that request from the source volume. However, if a read requests a previous version of the data, VSS maps that request to the appropriate block(s) and the data sent back will generally include blocks from both the source volume and the shadow copy volume.

Figure 1 shows where the VSS volume snapshot driver resides in the Windows Server 2003 layered driver model. For more information, consult the Windows Server 2003 online help that is available from the operating system's graphical user interface (GUI) or the Microsoft Developer Network (MSDN) at <http://msdn.microsoft.com>.

VSS also enables components associated with a backup system—such as business application software, storage management software, and storage hardware—to be used in an end-to-end process that helps ensure maximum backup efficiency. In this integrated approach, a writer (the business application software), a requester (the storage

management software), and a provider (the storage hardware driver) interoperate with VSS to provide a consistent, reliable backup snapshot of the application data. In the Figure 2 scenario, the writer is Exchange Server 2003, the requester is VERITAS Backup Exec™ storage management software, and the provider controls a Dell/EMC storage area network (SAN).

Exchange Server 2003: The writer

Exchange Server 2003 supports the Windows Server 2003 VSS feature, which helps reduce backup time and simplify backup management for Exchange data. The VSS writer service helps backup software obtain consistent, point-in-time shadow copies of data on a live Exchange server. When VSS receives a shadow copy request from backup software, VSS communicates with the running Exchange application (the writer) to pause new transactions, finish current transactions, and flush all the cached data to disk. VSS then communicates with the appropriate storage provider (in this scenario, the Dell/EMC storage array) to initiate the shadow copy process for the disk volumes that contain Exchange Server 2003 data. Once a shadow copy has been created, the backup software (in this scenario, VERITAS Backup Exec) then can copy data from the shadow copy to a tape without involving the Exchange application, thus reducing the impact of backup operations on Exchange Server 2003 performance and availability.

After the shadow copy has been created, VSS communicates with the Exchange writer service to signal that Exchange Server 2003 can resume writing to disk. The shadow copy process typically takes less than a minute. Clients using the Microsoft Outlook® 2003 messaging and collaboration client in cached mode most likely will not notice a disruption. Clients using earlier versions of Outlook may experience a delay of several minutes during the shadow copy process. Retaining the snapshot of an Exchange Server 2003

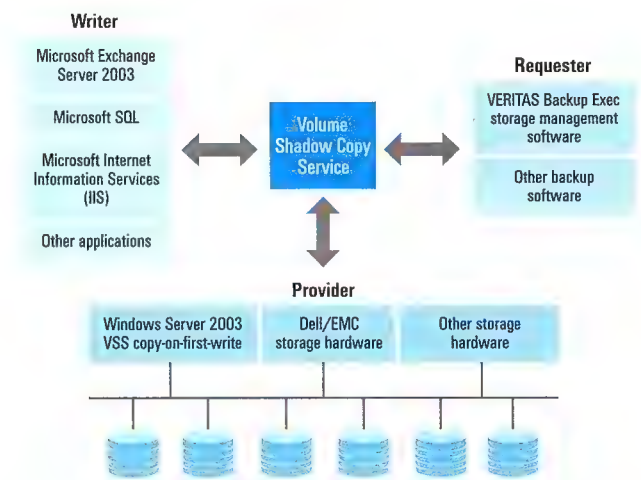


Figure 2. VSS-based backup configuration

	SnapView snapshot	SnapView clone
Description	Logical point-in-time copy using copy-on-first-write method	Block-level duplicate that allows forward and reverse synchronization
Storage requirement	One or more SnapView cache LUNs must be available; cache space is consumed when either the source LUN or the snapshot is modified	Clone LUN must be as large as the source LUN
Persistence	Optional	Yes
Required array software	EMC SnapView	EMC SnapView and EMC SnapView clones provider

Figure 3. Comparison of SnapView snapshots and clones

database on disk enables a faster restore operation, thereby helping to reduce the interruption to the Exchange service.¹

VERITAS Backup Exec: The requester

VERITAS Backup Exec for Windows Servers is designed to be a comprehensive backup tool for Microsoft Windows®-based server environments. Backup Exec 9.1 supports VSS-enabled backups, minimizing disruption of applications and services while data is being backed up. This enables Backup Exec to perform a nonintrusive backup by obtaining a consistent copy of data from VSS, which is done by temporarily stopping I/O to the data through the writer (Exchange, in this scenario) when a backup is initiated. Administrators can select the appropriate VSS writers from a list of supported shadow copy components during Backup Exec backup and restore operations. Backup Exec supports the following types of VSS writers:

- **Service state:** Critical operating system and application service data such as event logs, Windows Management Instrumentation (WMI), and Removable Storage Manager (RSM)
- **System state:** Critical operating system data such as Windows system files, Component Object Model+ (COM+) Class Registration database, registry, and Microsoft Active Directory® directory service
- **User data:** Microsoft SQL Server™, Exchange Server, Active Directory Application Mode (ADAM), third-party application and user data, and so on

Service state, system state, and user data compose the Backup Exec shadow copy components file system. Backup Exec supports only full backups of storage groups using the Exchange writer. To back up Exchange data, system administrators select the Exchange writer during the backup operation—the Backup Exec interface lists the available Exchange servers and information stores available for backup. In addition, administrators can use the Backup Exec Advanced Open File Option by selecting VSS as the service for open file operations.²

Dell/EMC storage array: The provider

EMC SnapView software and the SnapView clones provider are optional components available for Dell/EMC CX300, CX400, CX500, CX600, and CX700 storage arrays. SnapView snapshots are logical point-in-time copies of a logical storage unit (LUN), and SnapView clones are block-level duplicates of a LUN. Figure 3 compares SnapView snapshots and clones.

Administrators can create and manage snapshots and clones using the Dell/EMC storage array. The Dell/EMC VSS hardware provider—installed on a Dell™ PowerEdge™ server running Windows Server 2003—enables VSS to use EMC SnapView and the SnapView clones provider to create hardware-resident shadow copies on a Dell/EMC CX300, CX400, CX500, CX600, or CX700 array attached to the server. Figure 4 shows the hardware configuration for the Exchange backup scenario described in this article, including key software components that must be installed on both the PowerEdge server and the Dell/EMC storage array.

For the Dell/EMC hardware provider to register itself with VSS, the Windows Server 2003 Distributed Transaction Coordinator (DTC) service must be running when the provider is installed. After installing the provider, administrators must run `vssadmin list providers` to verify that the Dell/EMC hardware provider has been registered successfully. The hardware provider translates VSS API calls into EMC® Navisphere® command-line interface (CLI) statements that control the Dell/EMC hardware. The hardware provider can produce differential shadow copies by using `navicli.exe` to invoke

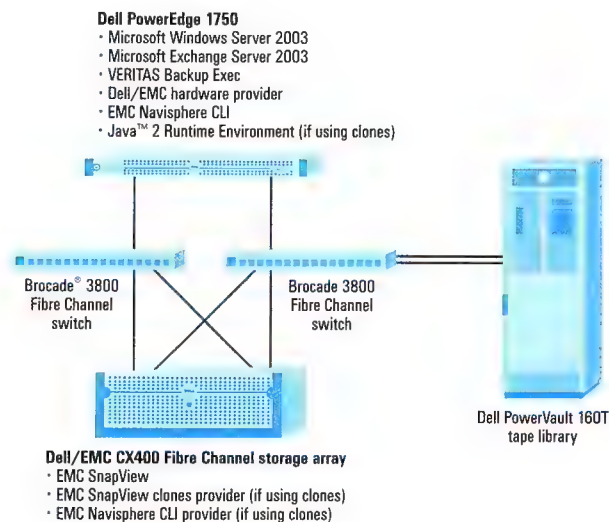


Figure 4. Integrated VSS-based backup configuration

¹ For more information, see the Exchange Server 2003 and Windows Server 2003 documentation online at <http://www.microsoft.com>.

² For more information, see the help and product documentation for Backup Exec software online at <http://www.veritas.com>.

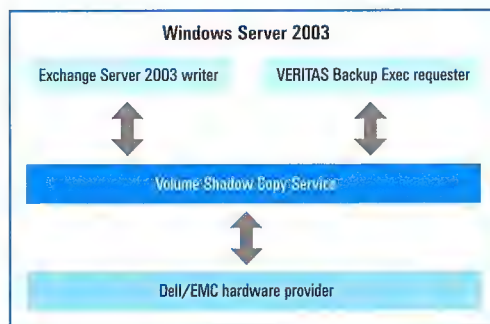


Figure 5. Application and storage management software for the VSS shadow copy backup scenario

SnapView and create a snapshot. Similarly, the hardware provider can create plex shadow copies, which are essentially duplicates of the source LUN, by using navicli.jar to invoke the SnapView clones provider to create a clone.³

Integrated VSS backup: Process flow

Figure 5 shows application and backup software used in the shadow copy-based backup scenario. In this example, a system administrator configures a backup schedule for Exchange Server 2003 through the Backup Exec software, specifying that the Exchange writer perform this service. When Backup Exec runs the scheduled job, the VSS requester informs the VSS service that Exchange Server 2003 storage groups will be affected. VSS then communicates with the Exchange writer to pause new transactions, finish current transactions, and flush all the cached data to disk.

VSS then communicates with the Dell/EMC hardware provider to create a hardware-resident shadow copy of the storage volumes containing the Exchange Server 2003 storage groups. The hardware

provider invokes EMC SnapView to create a snapshot for the requested disks in the EMC storage array and returns the snapshot to VSS, which mounts the snapshot for Backup Exec. Once this step is completed, VSS instructs Exchange Server 2003 to resume normal operations. Backup Exec backs up the Exchange Server 2003 data and logs from the disk snapshots to the tape backup devices. After the backup job is completed successfully, Backup Exec instructs VSS to delete the shadow copies

VSS can enable administrators to streamline the storage management process—reducing the time, complexity, and cost of online backups considerably.

created for backing up the storage group volumes. Finally, VSS instructs the Dell/EMC hardware provider to delete the snapshots requested for the backup operation.

A cost-effective, integrated approach to system backup

IT administrators can build reliable, integrated backup systems using standards-based enterprise components such as Dell PowerEdge servers, Dell/EMC storage arrays, Dell PowerVault™ tape libraries, and VERITAS Backup Exec software. The Volume Shadow Copy Service, included in Microsoft Windows Server 2003, enables administrators to create an integrated approach to system backup and recovery that helps safeguard business-critical data without disrupting high-availability services like customer databases, e-commerce applications, and corporate messaging systems. In addition, the integrated backup approach described in this article can simplify system administration and storage management, helping to reduce total cost of ownership for today's complex enterprise environments. ☛

Ananda Sankaran (ananda_sankaran@dell.com) is a systems engineer in the High-Availability Cluster Development Group at Dell. His current interests related to high-availability clustering include cluster management, databases, SANs, and tape backup. Ananda has a master's degree in Computer Science from Texas A&M University.

Kevin Guinn (kevin_guinn@dell.com) is a systems engineer in the High-Availability Cluster Development Group at Dell. His current interests include storage management and business continuity. Kevin has Microsoft Certified Systems Engineer (MCSE) and EMC Proven™ Professional certifications, and has a B.S. in Mechanical Engineering from The University of Texas at Austin.

Dat Nguyen (quocdat_nguyen@dell.com) is a systems engineer in the High-Availability Cluster Development Group at Dell. His responsibilities include developing SAN-based high-availability clustering products that comprise Dell servers and Dell/EMC Fibre Channel storage systems. Dat has a B.S. in Electrical Engineering from The University of Houston.

FOR MORE INFORMATION

Microsoft Volume Shadow Copy Service:

http://www.microsoft.com/technet/treeview/default.asp?url=/technet/prodtechnol/windowsserver2003/proddocs/techref/w2k3tr_vss_how.asp

VERITAS Backup Exec:

<http://www.veritas.com/products/category/ProductDetail.html?productId=bews>

Dell/EMC CX400 Fibre Channel storage array:

http://www.dell.cc.n/downloads/global/products/pvaul/en/cx400_spec.pdf

³ For more information, see the documentation for EMC SnapView and the Dell/EMC hardware provider online at <http://www.emc.com>.

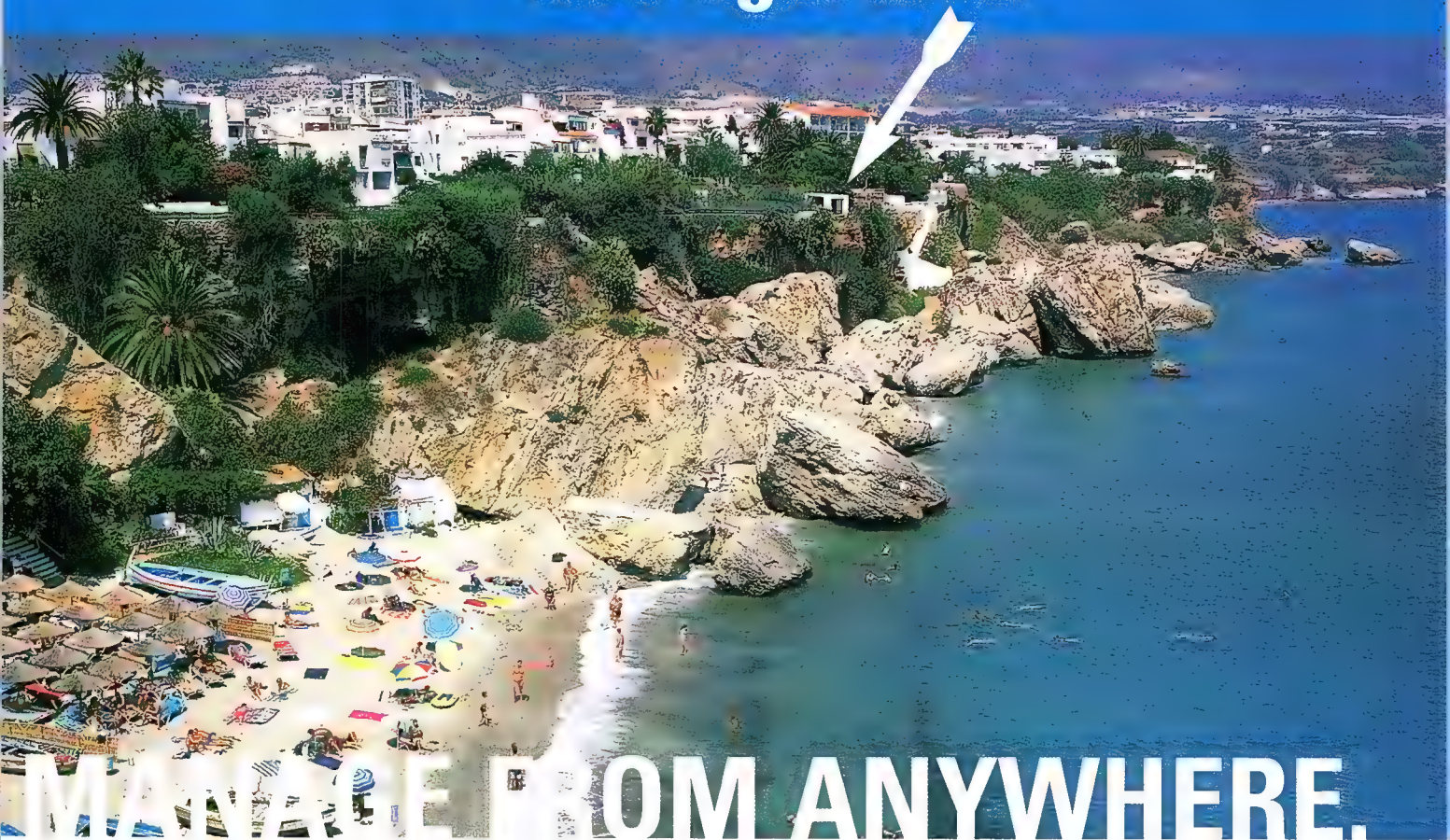
Dave is on vacation.

He's been notified of a problem with his company's servers.

He's able to manage those servers from anywhere.

He's not cutting his vacation short.

He's right here.



The Dell remote console switch provides the traditional functionality of KVM switching with the added benefit of **KVM Over IP™**. Using Dell's remote console software, simply point-and-click to take control of your servers.

Whether you're at the rack, in your office or across the globe your data center will always be within your reach.



Take **control** of your servers with the **Dell Remote Console Switch**.

Digital Availability. Easy as **DELL**

www.dell.com

Dell and the Dell logo are trademarks of Dell Inc. KVM OVER IP is a trademark of Avocent Corporation or its affiliates.

Managing Data Protection

with Red Hat Linux and Dell PowerVault Tape Autoloaders and Libraries

Protecting data from disaster is a critical concern for any enterprise. This article explores Dell™ PowerVault™ tape backup systems in Red Hat® Linux®-based environments, examining basic operations and configuration options. In addition, native Linux tools and third-party software that can help IT administrators implement an effective backup plan are addressed.

BY TESFAMARIAM MICHAEL AND RICHARD GOODWIN

Protecting business-critical data from disaster is one of the most important tasks that system administrators perform. Data backup systems usually involve tape drives, tape autoloaders, or tape libraries. Autoloaders are particularly well suited to small, networked computer environments. This article explores the basics of autoloader operations under the Linux® operating system (OS), including the device mappings of tape drives under Linux and commands to dump and restore data. It also provides a brief description of Dell™ PowerVault™ autoloaders, as well as the fully tested and validated backup software applications available from Dell.

Distinguishing between tape autoloaders and libraries

The terms *autoloader* and *library* often are used interchangeably, but the two device types differ slightly. An autoloader refers to a single drive unit with storage slots for multiple tapes; autoloaders can move tapes between the drive and the slots. The capacity of an autoloader generally is limited to between 8 and 20 tapes. In contrast, a tape library has capacity for multiple tape drives and a larger number of slots. Dell PowerVault tape libraries can range from 2 to 36 tape drives and 24 to 1,344 slots.

Figure 1 summarizes configuration options for Dell PowerVault autoloaders and libraries.

Like an autoloader, the library shuttles tapes between the tape drives and storage slots. In an autoloader, this process usually is performed by a sliding carousel or a caddy system; in a library, a robotic arm moves the tapes. To track its tape inventory, the autoloader or library either uses a bar code labeling system with a bar code reader or identifies a tape by the specific numeric slot in which it is located. The Linux OS and most production applications generally are not affected by the difference in how tape units operate. Consequently, when either *autoloader* or *library* is used in this article, it refers to both types of device.

Dell tape devices address various storage needs and are compatible with the Linux OS. In fact, Dell qualifies the entire PowerVault line of tape storage products for Linux—stand-alone tape drives as well as tape autoloaders and libraries. This article uses the Dell PowerVault 132T tape library as an example tape device because it can provide multiple tape drives and multiple slots, and can be controlled by Linux commands that are normally used with tape autoloaders. Although the information provided in this article can

apply to any Linux distribution, some details may be specific to the Red Hat® Linux OS.

Preparing the tape system for a Linux environment

A SCSI tape drive should be connected to a SCSI controller. For Dell PowerVault tape systems, this controller most likely will be an Adaptec® 39160 or 2940 for Low Voltage Differential (LVD) devices or Adaptec 3944 or 2944 for High Voltage Differential (HVD) devices. The Linux driver for these controllers, `aic7xxx`, must be loaded to access all the devices connected to a controller, such as the tape drive. Once the tape drive is connected and configured, the driver should detect the tape drive. To verify the detection of the tape drive, administrators should check for its entry in `/proc/scsi/scsi`. For more information about identifying PowerVault systems in `/proc/scsi/scsi`, see “Special considerations for LUN-based devices under Linux.”

The drivers required for Fibre Channel controllers will vary; administrators should consult the documentation provided with the controller. Once the appropriate driver is loaded, Fibre Channel-attached devices will register with the SCSI tape subsystem in the same manner as SCSI devices described in this article.

RPM packages for tape drive operation

Native Linux applications required to operate an autoloader include `tar`, `cpio`, `mt`, and `mtx`. Depending on the type of Linux installation, most if not all of these applications should already be installed on the system. Administrators can determine whether each of these packages is installed by querying the RPM™ (Red Hat Package Manager) database with `rpm -q rpm-name`, where `rpm-name` is the name of the application—`tar`, `cpio`, `mt`, or `mtx`. If this returns `rpm-name` with a version number, the package is installed. If a package is not installed, administrators can obtain it from the Red Hat Linux installation CDs or from <http://www.redhat.com>, and install it by entering `rpm -ivh application-rpm-name`.

Device mapping of SCSI tape system

In Linux, all devices are viewed as files with special attributes to the kernel, so applications can open, close, read, and write the files using system calls. In most Linux distributions, these device files (sometimes known as device nodes) are found in the `/dev` directory

Product	Maximum number of drives	Maximum number of slots	Bar code reader	Fibre Channel (FC)/SCSI
PowerVault 122T autoloader	1	8	Optional	SCSI only
PowerVault 132T tape library	2	24	Yes	FC or SCSI
PowerVault 136T tape library	6	72	Yes	FC or SCSI
PowerVault 160T tape library	12–36	264–1,344	Yes	FC only

Figure 1. Configuration options for Dell PowerVault tape automation devices

Device	Rewind node	No-rewind node
1st SCSI tape drive	<code>/dev/st0</code>	<code>/dev/nst0</code>
2nd SCSI tape drive	<code>/dev/st1</code>	<code>/dev/nst1</code>
<i>n</i> th SCSI tape drive	<code>/dev/st[n-1]</code>	<code>/dev/nst[n-1]</code>

Figure 2. Device nodes corresponding to physical devices

of the root file system. Device files are created in this directory with their respective attributes, such as major and minor numbers, device type (character or block), and permissions. The major number identifies the device type; the minor number informs the kernel about the special characteristics of the device. The following is an example of the device attributes of a SCSI tape drive:

```
crw-rw---- 1 root disk 9, 0 Jul 23 2003 /dev/st0
crw-rw---- 1 root disk 128, 0 Jul 23 2003 /dev/nst0
```

The device nodes for SCSI tape drives are `/dev/stX` and `/dev/nstX`, where `X` is an integer. When loaded, the `st` driver associates the tape drive with the device node depending on the order of detection. For example, the first SCSI tape drive detected will be assigned `/dev/st0` (see Figure 2). Two device nodes can be used when operating tape drives: rewind and no-rewind. The rewind node (`/dev/stX`) rewinds the tape to the beginning after every operation, whereas the no-rewind node (`/dev/nstX`) stops the tape wherever an operation leaves it, allowing multiple archives to be stored on a single tape.

Administrators should choose between rewind (`/dev/st0`) and no-rewind (`/dev/nst0`) devices. However, rewind devices can overwrite data because the tape is rewound after every operation. This makes it impossible to back up multiple archives to a single tape using the rewind device. For example, if an administrator enters the `mt -f /dev/st0 eod` command to prepare the tape for appending new data, the tape will be forwarded to the end of data (eod) of an archive position. Then the drive is closed and the tape is rewound to the beginning. At that point, if the system writes new data to the tape, it will overwrite the existing data instead of appending it to the tape as planned. This problem can be avoided simply by using the no-rewind node (`/dev/nst0`). Administrators should create a symbolic link from the no-rewind device to `/dev/tape` and use this link when operating the tape. To create this link, administrators should enter `ln -s /dev/nst0 /dev/tape`.

SCSI generic interface and tape drives

SCSI tape drives can be manipulated using the SCSI generic interface. This interface provides general access to SCSI devices from a user space application—an application that resides outside the kernel space. When the Linux SCSI Generic (`sg`) driver is loaded, each SCSI device in the system, detected by its respective driver, is

Attached devices:

```

Host: scsi0 Channel: 00 Id: 00 Lun: 00
  Vendor: DELL      Model: PERCRAID RAID5  Rev: V1.0
  Type:   Direct-Access      ANSI SCSI revision: 02
Host: scsi0 Channel: 00 Id: 01 Lun: 00
  Vendor: DELL      Model: PERCRAID Mirror Rev: V1.0
  Type:   Direct-Access      ANSI SCSI revision: 02
Host: scsi2 Channel: 00 Id: 05 Lun: 00
  Vendor: DELL      Model: PV-122T        Rev: D37r
  Type:   Medium Changer      ANSI SCSI revision: 02
Host: scsi2 Channel: 00 Id: 06 Lun: 00
  Vendor: HP        Model: Ultrium 1-SCSI Rev: E32K
  Type:   Sequential-Access     ANSI SCSI revision: 03

```

Figure 3. Example of SCSI devices listed in `/proc/scsi/scsi`

associated with the `/dev/sgX` device node, where *X* ranges from 0 to 256. The detection method of the `sg` driver is similar to that of a tape drive. For example, `/dev/sg1` and `/dev/sg2` are the first and second SCSI devices detected, respectively.

The `/proc/scsi/scsi` interface of the virtual `proc` file system presents a list of all the SCSI devices currently detected in a system. Similarly, the `sg` driver provides a list of these SCSI devices in `/proc/scsi/sg/device_strs`. Figure 3 shows an example of the `/proc/scsi/scsi` file in a system with several SCSI devices.

As shown in Figures 3 and 4, a one-to-one mapping exists between the devices listed in `/proc/scsi/scsi` and `/proc/scsi/sg/devices_strs`. The first two `sg` devices (`/dev/sg0` and `/dev/sg1`) are assigned to the PERCRAID RAID volumes. The tape changer is mapped to `/dev/sg2`, and the tape drive is mapped to `/dev/sg3`.

Once the `sg` device node of the tape changer is known, a symbolic link from the device to `/dev/changer` should be created. This link helps simplify administration by eliminating the need to remember the device node name every time the device needs to be accessed. The link can be used when manipulating the tape changer with the `mtx` program, which uses this device by default. In Figure 3, because the third device in the list is the tape changer, `/dev/sg2` is the device for it and can be linked to `/dev/changer`. This link can be created by issuing the `ln -s /dev/sg2 /dev/changer` command.

Using native Linux commands to perform tape backups

Several programs under Linux—such as `tar`, `cpio`, `dd`, `dump` and `restore`—can be used for backups. In addition, the `mt` and `mtx` tools allow administrators to perform necessary autoloader and media operations such as load/unload and forward/rewind. This section provides a brief description of some of these programs; for additional information about these programs, please see their respective man pages. The example scenarios in this section assume the SCSI tape drive contains a blank tape and the system has a `/backup` directory.

```

DELL    PERCRAID RAID5    V1.0
DELL    PERCRAID Mirror   V1.0
DELL    PV-122T           D37r
HP      Ultrium 1-SCSI    E32K

```

Figure 4. Example of SCSI devices listed in `/proc/scsi/sg/device_strs`

Tape archive. The `tar` archiving program stores to and extracts from an archive in a tape drive or a normal file. The common syntax of `tar` is as follows:

```

tar -mode -options [archive-device
or archive-name] [files-to-archive]

```

In this syntax, *mode* can be `-c` for create (backup), `-x` for extract (restore), or `-t` for table of contents (list). The variable *options* can include `-v` for verbose, `-f` for archive destination (in create mode) or source (in extract or table of contents mode), or a combination thereof. For example, to archive the `/home` directory to a no-rewind tape, the following syntax would be used (assuming the symbolic link of `/dev/tape` has been created):

```
tar -cvf /dev/tape /home
```

To extract this archive from the tape, administrators should position the tape at the beginning of the archive, and then extract the archive to the appropriate directory. This operation can be performed as follows:

```

cd /backup
mt asf 0
tar -xvf /dev/tape

```

Copy input/output. The `cpio` program moves data to and from an archive, and also works well for backups. Unlike `tar`, `cpio` reads the name of the file it is to process from standard input. The common syntax of `cpio` is as follows:

```
cpio -mode -options <file-name-list [>archive-name]
```

In this syntax, *mode* can be `-o` for creating an archive, `-i` for extracting an archive, and `-t` for listing a table of contents for an archive. The variable *options* can include `-d` for directory creation, `-m` for preserve-modification time, `-u` for unconditionally replacing files, and `-v` for verbose. A common method for generating a list of files for `cpio` is to use commands such as `find`, which sends its output to `cpio`. The `find` command has some features that make it useful

**They're looking to you to solve the problem.
Look to Dell to teach you how.**



Dell™ Training & Certification

How can you realize the potential and maximize the value of your organization's technology assets? With Dell Training & Certification. Dell makes it simple, recognizing participants' problems and providing the resources and knowledge to overcome them. Through comprehensive and affordable online training, instructor-led courses and certification exams, Dell Certification Programs deliver the expertise required to install, configure and manage Dell server, storage and networking solutions. That includes Dell/EMC® storage area networks, Dell PowerConnect™ networks, Dell PowerEdge™ servers and more.

If they're turning to you for answers, turn to Dell for training. To learn more, enroll or get a copy of the latest *Dell Power Solutions* technical journal, visit www.dell.com/training/lookingtoyou.

Certification made easy. Easy as **DELL™**

Call 1-866-360-3506 Click www.dell.com/training/lookingtoyou

when performing full or incremental backups. The tar example used earlier can be performed using cpio:

```
find /home | cpio -o > /dev/tape
```

This archive can be extracted as follows:

```
cd /backup
mt asf 0
cpio -idv --no-absolute-filenames < /dev/tape
```

Magnetic tape. The mt program controls magnetic tape drive operations. This tool can be used to place the tape at a certain position, rewind and forward, eject, erase, check drive status, and so on. The common syntax of mt is as follows:

```
mt [-f device] operation [count] [argument]
```

In this syntax, *device* is the device node of the tape drive, which is /dev/tape in the previous example; *operation* can include status, eject, load, offline, and so forth. The *argument* is specific to the issued operation. For example, to report the status of the drive, administrators would issue the following command:

```
mt -f /dev/tape status
```

Because a symbolic link was created, this task also can be performed using mt status. From the message the status command generates, the value of *File Name* represents the number for archives in the tape up to the current position; and the value of *Block Number* represents the number of blocks in the archive if the tape is positioned at the end of an archive (if it is positioned at the beginning of an archive, the value is zero).

Media changer. The mtx program manipulates SCSI media changer devices such as tape autoloaders.¹ This program includes commands for loading and unloading a tape to and from the drive, providing inventory of all the slots, reporting all drives and media in the device, and so forth. The common syntax of mtx is as follows:

```
mtx [-f scsi-generic-device] [...] command
```

In this syntax, *scsi-generic-device* is /dev/sgX and X is the detection order of the changer. The variable *command* can include load/unload, inquiry, status, first, last, and next. The following example is a command that unloads media from the drive to the first slot:

```
mtx -f /dev/sg2 unload 1
```

Because the /dev/changer symbolic link was created earlier, /dev/changer can be used in place of /dev/sg2 or the -f /dev/sg2 arguments can be omitted altogether.

Examining the three types of backup

Building upon the basics that have been presented in this article, administrators can write a complete backup script. This script can be added to the cron job of the backup server to automate the backup process.²

Backups can be divided into three types: full, incremental, and differential. Administrators should evaluate the benefits of each type carefully when determining a backup plan. Often, a combination of these types is necessary—for example, full backups performed weekly and differential or incremental backups performed daily. The type of backup, the data to back up, and the frequency of backups should be dictated by the criticality of the data. Regardless, regular backups are necessary and must be performed.

Full backup. A full backup writes every file in a system to backup media. Because a full backup is costly in both time and media, this approach is not efficient when only a few files have changed since the last backup.

Incremental backup. An incremental backup includes only files that have changed since the last backup of any kind. Before writing a file to the backup media, the backup software identifies the modification time of that file. If the modification time is more recent than the last backup time, then the file is backed up. If a file has changed more than once since the last full backup, the most recent file will not replace the version(s) of the file already backed up—instead, the newest version will also be backed up so that each revision of the file is saved. This type of backup requires less time and backup space, but should be used in combination with full backups. For example, administrators commonly implement full backups weekly and incremental backups daily.

Differential backup. In a differential backup, all files that have been modified since the last full backup are backed up on every subsequent, or *differential*, backup until the next full backup. Unlike incremental backups, only the most recent version of the file is saved—the previous version of the file is overwritten. This approach reduces the restore process time by allowing administrators to restore all data using only the most recent full backup media and the most recent differential backup media.

A combination of the tools discussed in the previous section can provide all the functions required for a complete backup. However, administrators must write some scripts and thoroughly test those scripts to help ensure a reliable software backup system.

¹ Please note that mtx was removed from the initial release of Red Hat Enterprise Linux 3, so administrators using this OS must download mtx from <http://www.redhat.com> or the mtx site at <http://mtx.badtux.net>.

² See the crontab and cron man pages for information on how to set up cron jobs. These man pages can be viewed by issuing `man crontab` and `man cron`, respectively.

If script development is not an option, administrators may consider several third-party backup software tools. Three third-party tools that Dell offers are discussed in the next section.

Exploring Linux backup software

Red Hat Linux, like most Linux distributions, offers several customizable backup applications such as AMANDA and taper. Although tar, mtm, and other freely available Linux utilities can

provide excellent control over tape libraries and drives, administrators often need a flexible approach that can work out of the box with minimal configuration.

AMANDA. Popular open source utilities such as tar and mtm fully support autoloaders and libraries, but system administrators may require tools that are more configurable or capable of being automated. The AMANDA program is an open source tool that can meet such needs for flexibility and automation. AMANDA, which

SPECIAL CONSIDERATIONS FOR LUN-BASED DEVICES UNDER LINUX

To verify the detection of a tape drive, administrators should check for its entry in `/proc/scsi/scsi`. Current versions of Linux may not scan the logical storage unit (LUN) ID of every device. This can result in some PowerVault devices not being identified or listed in the `/proc/scsi/scsi` output. Administrators can follow these steps to enable support for such devices.

1. Type `cat /proc/scsi/scsi`. The output will look similar to the following:

```
Attached devices:
Host: scsi2 Channel: 00 Id: 00 Lun: 00
Vendor: DELL Model: PERCRAID Stripe Rev: V1.0
Type: Direct-Access ANSI SCSI revision: 02
Host: scsi3 Channel: 00 Id: 00 Lun: 00
Vendor: DELL Model: PV-136T-FC Rev: 4193
Type: Unknown ANSI SCSI revision: 03
```

2. Identify the host adapter, channel number, target ID number, and LUN number for the first LUN of the device to be configured. In this example, the PowerVault 136T-FC (Fibre Channel interface) is shown at the address, or *nexus*, 3 0 0 0—which means host adapter 3, channel number 0, ID 0, and LUN 0.
3. Determine the additional LUN IDs that are configured on the PowerVault device. Refer to the PowerVault documentation for instructions on determining the LUN configuration. One exception is the PowerVault 122T-VS80, which always has the tape drive at LUN 0 and the robot at LUN 1.
4. For each additional LUN that needs to be discovered by Linux, issue the following command:

```
echo "scsi-add-single-device H C I L" >
    /proc/scsi/scsi
```

H C I L refers to the nexus described in step 2. So, if the PowerVault 136T robot was configured at LUN 1, type:

```
echo "scsi-add-single-device 3 0 0 1" >
    /proc/scsi/scsi
```

Repeat this step for each additional LUN. The echo command will force a scan of each device at the given nexus.

5. Type `cat /proc/scsi/scsi` again to verify that all devices are now listed. The output will look similar to the following:

```
Attached devices:
Host: scsi2 Channel: 00 Id: 00 Lun: 00
Vendor: DELL Model: PERCRAID Stripe Rev: V1.0
Type: Direct-Access ANSI SCSI revision: 02
Host: scsi3 Channel: 00 Id: 00 Lun: 00
Vendor: DELL Model: PV-136T-FC Rev: 4193
Type: Unknown ANSI SCSI revision: 03
Host: scsi3 Channel: 00 Id: 00 Lun: 01
Vendor: DELL Model: PV-136T Rev: 2.88
Type: Medium Changer ANSI SCSI revision: 02
Host: scsi3 Channel: 00 Id: 00 Lun: 02
Vendor: QUANTUM Model: SDLT320 Rev: 4646
Type: Sequential-Access ANSI SCSI revision: 02
Host: scsi3 Channel: 00 Id: 00 Lun: 03
Vendor: QUANTUM Model: SDLT320 Rev: 4646
Type: Sequential-Access ANSI SCSI revision: 02
```

Administrators should add the echo command to the Linux boot scripts because the device information is not persistent and must be created each time the system boots up. One example file that can be used for storing the commands is `/etc/rc.local`. Note that configuring additional devices on a server or a storage area network (SAN) can cause the devices to be reordered, which requires administrators to modify the commands. If the Fibre Channel adapter supports Persistent Bindings or an equivalent function, it can be enabled to reduce the chance of devices being reordered upon discovery.

stands for Automated Maryland Automatic Network Disk Archiver, was designed to collect data from distributed clients, servers, or both, and to send that data to a single master server. The data is first spooled to disk and then offloaded to a high-capacity tape drive. Additional modules and scripts support the use of autoloaders and libraries, including support for devices that have bar code readers.

To install and configure AMANDA, administrators must build the packages from source (or locate a compatible package for the Linux distribution being used) and configure text files that will dictate which clients to back up, what data to back up, and how to move the data to tape. Newer versions of AMANDA also support the use of Samba, a tool that enables a UNIX®-based server to act as a file server for Windows clients. For a full description of the installation and configuration of AMANDA, including FAQs and examples, visit the AMANDA home page at <http://www.amanda.org>.

Yosemite Technologies® TapeWare®. This unique cross-platform tool is available as a qualified Dell package when used with Dell PowerVault tape libraries, autoloaders, and drives in direct attach SCSI mode. TapeWare provides an easy-to-use graphical user interface (GUI), as well as a character interface that is identical across the platforms it supports. Dell tape devices support TapeWare running on Microsoft® Windows®, Novell® NetWare®, and Red Hat Linux operating systems.³

Administrators can choose from either the X Window System (Qt) GUI or the ncurses-based character interface to set up and manage backup and recovery activities. TapeWare provides complete scheduling support for rotating backups, days of the week, and many other options. Dell has qualified the complete line of Dell PowerVault tape products, including the PowerVault 136T, for use with TapeWare.³ Wizards are available to guide administrators through most processes, making this software appropriate for those who are inexperienced with tape backup activities. All device functions—including importing and exporting of tapes, media management, and device status information—are supported and available through both user interfaces.

VERITAS NetBackup™. Another backup tool available for enterprise environments is VERITAS NetBackup software. Dell has qualified the NetBackup application to run across its entire range of PowerVault tape products, including the PowerVault 160T enterprise library. In addition to direct attach SCSI hardware, NetBackup also supports PowerVault libraries attached to storage area networks (SANs). Currently Dell supports NetBackup for Linux under Red Hat Linux Advanced Server 2.1. NetBackup also is available for Windows, NetWare, and all major UNIX operating systems.

NetBackup uses the concepts of *master server* and *media server* to distribute tape storage duties across multiple-server systems.

The master server schedules data protection operations, stores the centralized database, and manages other administrative functions. Media servers handle the actual data transfer to tape hardware. By default, a master server is also a media server, so a single server can completely manage the tape hardware—or the master server can distribute duties to other media servers. NetBackup supports Red Hat Advanced Server 2.1 as both a master server and a media server.

NetBackup includes a Java™-based GUI as well as extensive command-line utilities for management. The powerful command-line interface offers administrators unique abilities to script and manage NetBackup in ways not available through the GUI. This capability underscores one of the major features of NetBackup: the flexibility to extend and customize the software to meet specific enterprise needs.

Backing up the future of Linux

Linux has grown from its roots as a hobby-level OS to an enterprise-class platform that can be suitable for many commercial applications. As the use of Linux increases, the amount of data that will need to be protected also will increase. Backup systems are available for almost any budget and need, and can be implemented by both experienced system administrators and those new to the Linux OS. ☞

Tesfamariam Michael (tesfamariam_michael@dell.com) is a software engineer on the Linux Development Team of the Dell Product Group, which tests Linux on all Dell PowerEdge servers. Tesfamariam has an M.S. in Computer Science from Clark Atlanta University, a B.S. in Electrical Engineering from the Georgia Institute of Technology, and a B.S. in Mathematics from Clark Atlanta University. His areas of interest include operating systems and I/O devices.

Richard Goodwin (richard_goodwin@dell.com) is a software engineer on the Tape Hardware and Software Development team of the Dell Product Group. This team is responsible for the development and qualification of backup solutions for sectors ranging from the small to medium-size business (SMB) market to enterprise customers. Richard's areas of interest include storage systems and digital video technologies.

FOR MORE INFORMATION

Dell PowerVault tape backup storage:

<http://www1.us.dell.com/content/products/compare.aspx/tapeb?c=us&cs=55561=en&s=biz>

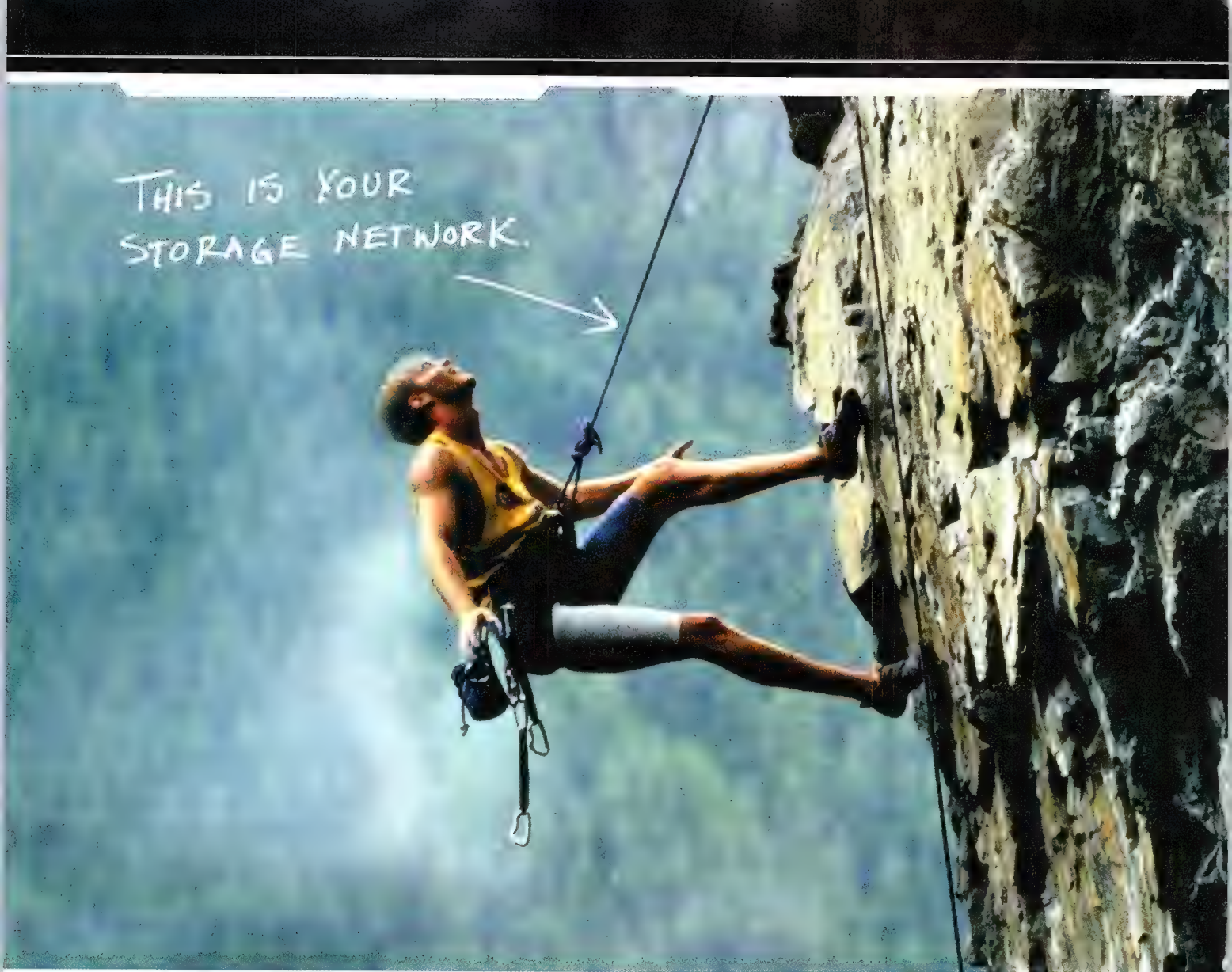
Dell and Red Hat Linux:

<http://www.redhat.com/onesource>

AMANDA:

<http://www.amanda.org>

³ At press time, TapeWare did not support SAN-attached Dell PowerVault tape devices.

A man in a yellow tank top and blue shorts is rappelling down a steep, craggy rock face. He is looking up towards the top of the cliff. A rope is attached to his harness and extends upwards. In the upper left corner, the text "THIS IS YOUR STORAGE NETWORK." is written in a white, chalk-like font. A white arrow points from this text towards the man's harness.

THIS IS YOUR
STORAGE NETWORK.

Will yours be there when you need it?

Keeping mission-critical data and applications available is of vital importance. And for companies of all sizes, there's no better lifeline than McDATA® multi-capable storage network solutions™. That's because these powerful solutions combine industry-leading hardware, software and services to deliver the scalability, reliability and investment protection that organizations like yours depend on. Just ask more than 80 percent of Fortune 100 companies that rely on McDATA to network the world's business data™.

Learn how you can benefit today from a storage services infrastructure engineered to make the on-demand computing environment a reality. To get your FREE "Business Advantages of a Real-time Storage Services Infrastructure" white paper, visit www.mcddata.com today.



McDATA™

Networking the world's business data™

Protecting Business-Critical Data

at Remote Offices

Remote office data protection can be a challenge for IT organizations that lack qualified backup administrators or adequate hardware budgets. However, as enterprises expand remote office operations, they cannot risk leaving business-critical data unprotected. Products such as VERITAS Backup Exec™ *for Windows Servers* and VERITAS Storage Replicator™ software can help make remote office data protection more efficient—and affordable.

BY SHERI ATWOOD AND MICHAEL PARKER

Many organizations invest heavily in protecting business-critical data at headquarters while leaving remote offices less protected—or not protected at all. However, today's business climate is driving enterprises to institute broader guidelines that protect their assets and minimize the risk of data loss. In response, IT administrators must overcome the challenges of limited staffing and capital budgets to ensure protection of critical business data, meet strict service level agreements, and conform to tough local and federal government regulations—or face potential data loss, revenue loss, or fines.

To protect data and help reduce the cost of backup operations at remote locations that have no on-site administrators, IT organizations may implement one of two strategies for maintaining, monitoring, and troubleshooting backup jobs. First, if tape drives already exist at remote locations, administrators can simplify and automate local storage management for distributed servers using products such as VERITAS Backup Exec™ *for Windows Servers* software and its Admin Plus Pack Option. Alternatively, organizations that do not need to perform the backup process at each remote location can protect data using products such as VERITAS Storage Replicator™ file-based

replication software, which efficiently copies data to a backup server at a centralized location.

Simplifying remote server deployment

Deploying and configuring backup servers can be a time-consuming, labor-intensive process for IT administrators. The Admin Plus Pack Option for VERITAS Backup Exec can help administrators reduce remote backup costs by enabling Backup Exec software—and all desired agents, options, and settings—to be installed efficiently on remote servers using the following methods:

- **Manual push installation:** Installs unique configurations from a centralized backup server to a single remote backup server over network connections.
- **Remote installation with cloned local settings:** Duplicates the configuration and settings that are installed on a remote server running Backup Exec software, from which the push operation is performed. This method mirrors the source backup server so that whatever agents, options, and settings are installed on the remote source also are installed on the cloned backup server.

- **Cloned image push installation:** Pushes a previously created cloned image of a unique Backup Exec configuration to another server. This method helps IT administrators improve efficiency when rolling out several similar backup servers.

- **Local silent installation:** Performs backups using a CD image created on a server that contains the desired Backup Exec configuration and settings. This CD image enables administrators to perform an automated, or *silent*, installation that does not prompt for any local user input. Although this method requires on-site deployment, it simplifies the process and reduces opportunity for error.

The ability to create and distribute jobs on remote servers enables organizations to improve efficiency without requiring on-site IT administrators.

The ExecView console also enables administrators to manage by exception, so that only errors and failed jobs appear in a particular view or are clearly identified in red. When managing several Backup Exec servers remotely, administrators can improve efficiency by zeroing in on critical errors or exceptions that require review or immediate action.

To assist in the management of remote and branch offices, the Admin Plus Pack Option offers advanced reporting that includes active alerts, alert history, configuration settings, device summary, event logs, media vault contents, and robotic library inventory. Reports can be viewed and printed in HTML format and distributed through e-mail. The capabilities of the Admin Plus Pack Option help IT administrators to track their protection and recovery services and to bill back individual departments on a regular basis.

Automating real-time data replication

Traditionally, organizations protect remote office data by deploying tape drives, tape media, and backup software—and by hiring administrators at each location to manage the tape backup process. When qualified backup administrators are not available, organizations may assign untrained employees the task of administering daily tape backup and restore operations. However, relying on untrained administrators to perform backups can increase an organization's chance for data loss because failed backups may go unnoticed.

An alternative approach to protecting remote office data enables administrators to combine data replication with traditional

Beyond the logistical challenges administrators must address when deploying remote servers, the development and setup of backup jobs can be extremely time-consuming. If several backup servers perform similar functions, the Admin Plus Pack Option for VERITAS Backup Exec can help administrators streamline job creation and distribution for remote servers. Because jobs, job templates, and selection lists can be copied between servers running Backup Exec software, administrators can create these components on one backup server and then copy them to one or more remote backup servers on the network (see Figure 1). The ability to create and distribute jobs on remote servers enables organizations to improve efficiency without requiring on-site IT administrators.

Managing remote office backups

Once administrators have completed backup server deployment and job setup, they can begin backup operations. VERITAS Backup Exec ExecView™ software, which is included in the Backup Exec package, provides a Web-based console that enables administrators to monitor jobs, devices, and alerts on hundreds of local or remote servers running Backup Exec software. Management functions include the following:

- Monitoring any backup server's active, scheduled, and completed jobs
- Pausing and resuming media servers and devices
- Starting scheduled jobs and canceling jobs
- Creating backup server groups
- Receiving and responding to alerts
- Receiving e-mail or pager notifications of events

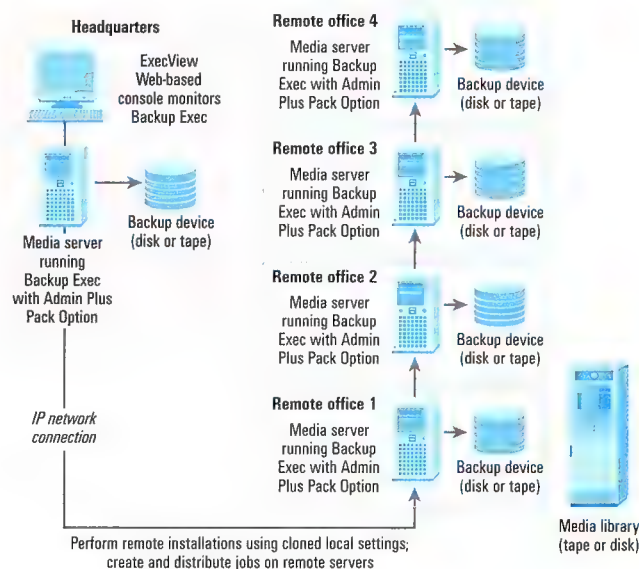


Figure 1. Managing remote or distributed servers

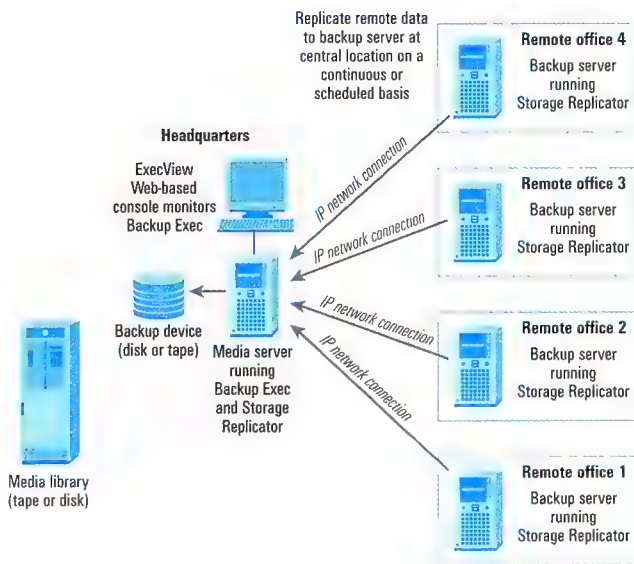


Figure 2. Replicating data back to headquarters


backup policies. For example, VERITAS Storage Replicator for Microsoft® Windows NT®, Windows® 2000, Windows Server™ 2003, and Windows Storage Server 2003 operating systems can help administrators efficiently protect their remote office data by replicating data to a backup server at a centralized location. This approach also helps IT organizations reduce costs by eliminating tape drives, tape media, and backup administrators at remote locations. VERITAS Storage Replicator can support hundreds of nodes and replication processes from one centralized console. Administrators simply push Storage Replicator out from the central location to remote offices and manage the remote offices from the central console.

The data replication approach discussed in this article requires an IP network connection between the remote offices and the central location. In the scenario shown in Figure 2, VERITAS Storage Replicator is installed on all remote office servers as well as on the central server. In addition, VERITAS Backup Exec for Windows Servers software is installed at the central site to perform the backup to tape or disk at that location.

Replication can be performed on all files, including open files, continuously or on a scheduled basis. For continuous replication, Storage Replicator copies a changed block back to the central location every time data is written to a file at the remote office. Alternatively, to maximize network connections for business-critical production servers, administrators can schedule data replication only during off-peak hours. In addition, administrators can restrict the amount of bandwidth that is used for replication to preserve adequate network throughput for production servers.

Once the remote office data arrives at the central location over an IP connection, backup administrators can perform standard backups consistently for all enterprise data and can manage replication jobs at remote offices from the central location. When necessary, central backup administrators can restore remote data over the network without requiring remote office staff to manage and restore the data from tape.

Maximizing resources while reducing costs

Many administrators must strike a balance between implementing cost-saving measures and protecting business-critical data. Backup and data replication packages such as VERITAS Backup Exec for Windows Servers and VERITAS Storage Replicator offer administrators tools to simplify on-site backups and streamline remote office backups. Such tools help administrators reduce the cost of data protection and improve the efficiency of remote office backups through centralized management. 

Sheri Atwood (sheri.atwood@veritas.com) is a senior product manager in the VERITAS Business Continuity Group for replication and disaster recovery solutions.

Michael Parker (michael.parker@veritas.com) is a product marketing manager in the VERITAS Business Continuity Group for Windows data protection solutions. He has a degree in Economics from Northwestern University.

VERITAS Software Corporation (<http://www.veritas.com>) is a leading storage software company, providing data protection, application performance, storage management, high availability, and disaster recovery software.

FOR MORE INFORMATION

VERITAS Backup Exec for Windows Servers:
<http://www.veritas.com/backupexec>

VERITAS Storage Replicator:
<http://www.veritas.com/products/category/ProductDetail.html?productid=storagereplicator>

VERITAS VISION Utility Computing Conference, May 3–7, 2004:
<http://www.veritas.com/vision>

The Admin Plus Pack Option offers advanced reporting that includes active alerts, alert history, configuration settings, device summary, event logs, media vault contents, and robotic library inventory.

Streamlining Backup and Recovery Operations

Using Disk-based Protection

IT administrators are straining to protect massive amounts of data in the face of ever-shrinking backup windows. This article examines how VERITAS Backup Exec™ 9.1 *for Windows Servers* and VERITAS NetBackup™ 5.0 software can enable IT organizations to implement a disk-based data protection strategy that helps improve backup and recovery times while increasing system availability.

BY SCOTT KOSCIUK, MICHAEL PARKER, AND MARK THOMASON

To help meet stringent system availability requirements, administrators now can enhance traditional tape-only backup operations using fast, flexible disk-based backup and recovery techniques that do not encroach on business-critical applications. This approach can enable IT organizations to better maintain service level agreements (SLAs) and to help create more responsive, cost-effective data protection and disaster recovery strategies.

A high-performance data protection strategy integrates both disk and tape storage with an optimized backup application that can streamline backup and recovery operations. Although disk-based data protection is not likely to replace tape drives and tape robotics completely, it can enable more efficient backups and recoveries. For example, administrators can perform backups quickly from a primary disk to a backup disk and then copy the data from the backup disk to tape for long-term or off-site storage. In addition, the ability to restore data from snapshots that reside on the primary or backup disk can provide near-instantaneous data recovery.

Comparing disk-based to tape-based data protection

Although the throughput and capacity of tape devices has become competitive with that of disk drives over the past few years, tape is still a sequential-access medium and as such can be inflexible and cumbersome compared to disk media. Moreover, disk-based storage avoids mechanical delays that are inherent in tape libraries or devices, such as tape mounting, positioning, and availability. Given the fast random-access read performance of disk volumes—especially RAID volumes—administrators can achieve near-instantaneous disk backups and restores by leveraging *snapshots*, which provide a point-in-time image of a client's data on local or remote disk storage.

Using disk-resident snapshots, tape backups, or both, administrators can reduce network backup windows and free host CPU and I/O cycles to process business-critical applications. VERITAS® software enables administrators to back up data from the snapshot image, instead of directly from the client's primary data—thereby allowing client operations and user access to continue without

interruption during the backup. In addition, fast disk-based reads and writes allow administrators to schedule more frequent backups, which help to improve data protection by lowering the incidence of data loss.

When used as backup devices, disks can support simultaneous backup, recovery, and duplication operations. To back up multiple sources to a single tape drive, administrators traditionally use a multiplexing, or *interleaving*, approach. Multiplexing consolidates multiple streams into one stream while writing to a tape drive. This approach keeps the tape drive spinning, rather than starting and stopping to wait for additional data. Although multiplexing can greatly increase tape device efficiency, its main disadvantage can be slow recovery operations. Disk-based storage eliminates the need for multiplexing because disks are inherently random access devices.

Given the benefits of disk-based storage, tape may appear to have little future in data protection. However, tape still provides the best medium for long-term and off-site storage, which can make tape an important consideration for disaster recovery and business continuance planning. In addition, tape technology offers benefits that disk media has not yet achieved, such as:

- **Greater durability:** Tapes can survive large drops and tolerate rough handling better than disk drives, which need to be carefully packed for shipping.
- **Lower unit cost:** New large-capacity tapes are bigger and less expensive than today's disk drives, providing a lower cost per megabyte.

Understanding disk-based backup and recovery methods

Five methods of disk-based data protection are currently available: backup to disk, disk staging, inline copy, synthetic backup, and instant recovery (see Figure 1).

Backup to disk. The backup-to-disk approach writes the same data to a file on a disk volume as it would to a file on a tape volume. Therefore, when a backup-to-disk operation completes, a single file the size of the backup will exist on the target volume that contains all the files that were backed up. Products such as VERITAS Backup Exec™ 9.1 for Windows Servers and VERITAS NetBackup™ 5.0 software provide administrators with a graphical user interface (GUI) from which to configure backup-to-disk

Data protection application	Backup to disk	Disk staging	Inline copy	Synthetic backup	Instant recovery
VERITAS NetBackup 5.0	Yes	Yes	Yes	Yes	Yes
VERITAS Backup Exec 9.1 for Windows Servers	Yes	Yes	No	No	No

Figure 1. Support for disk-based data protection strategies

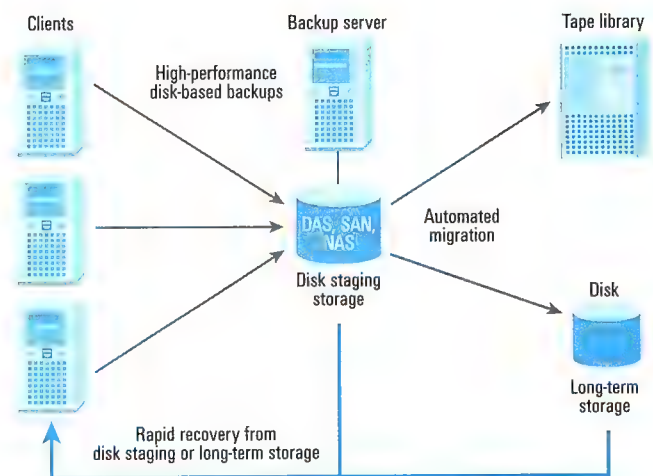


Figure 2. The disk staging process

policies and storage designations, and to perform full, archival backups as well as incremental or differential backups. Both products work with network attached storage (NAS), storage area networks (SANs), and direct attach storage (DAS).

Disk staging. The disk staging method writes backup data to a disk cache before sending the data to its final destination, disk or tape (see Figure 2). The purpose of disk staging is to use all available media to best advantage. For example, when staging a backup, administrators first copy the target data onto the disk cache and later move the backup image to tape according to the established disk staging schedule. Disk staging enables administrators to complete backups faster, shortening the backup window and thereby affecting business applications less than a direct backup-to-tape method.

The backup data remains in the disk staging storage unit until either the backup expires, based on the administrator-specified retention period, or another backup needs space in the disk staging location. When a backup application such as VERITAS NetBackup software detects a full disk staging location, it pauses the backup process, finds the oldest backup image that has been copied successfully to the final long-term storage destination, and deletes that data from the staging storage unit. For rapid restores, NetBackup can retrieve undeleted data from the disk staging storage unit before requesting the data from the secondary, long-term storage location.

Inline copy. The inline copy method writes backup data simultaneously to multiple destinations, such as disk and tape. Before the inline copy approach was introduced, administrators were required to duplicate backup application data as a secondary process, after they finished the initial backup from the client. By combining the backup and duplication operations, the inline copy approach can enable administrators to make an organization's electronic

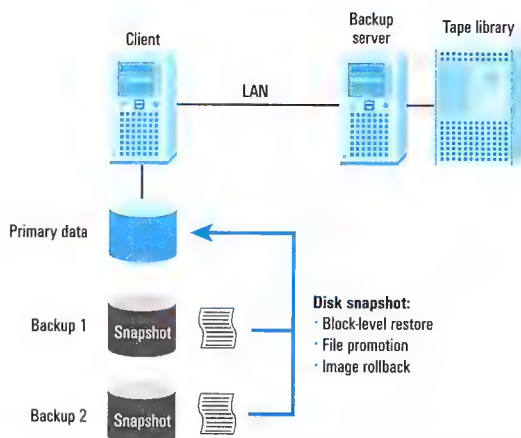


Figure 3. The instant recovery process

data-vaulting procedures more flexible and efficient, conserving time and IT resources.

Synthetic backup. When system availability requirements do not allow enough time to complete a full backup, administrators can create a synthetic backup that is identical to a current full backup. A synthetic backup assembles data from the system's previous full backup and subsequent incremental backups without involving network resources—the synthetic backup is completed on the media server. The ability to create synthetic backups can help administrators meet regulatory or SLA requirements for frequent full backups when performing actual archival backups over the network is too expensive or resource-intensive.

To create a synthetic backup, selected data is copied from the initial full backup and subsequent incremental backups to disk or tape as a single backup image. To construct a synthetic backup optimized for high-performance recovery operations, the VERITAS NetBackup software tracks not only which files were present on the volume for each backup, but also which data was moved or deleted

since the last full backup. A disk-based approach can dramatically increase the usability and efficiency of synthetic backups. For example, using disk storage for smaller, more frequent incremental backups can help administrators reduce, if not eliminate, the requirement to mount a large number of tapes.

Instant recovery. The instant recovery approach helps deliver the benefits of disk-based data protection without the need to perform backups or recoveries over the network (see Figure 3). By restoring

data from snapshots residing on a local disk, administrators quickly can resolve application corruption and end-user errors such as accidental deletions and overwrites. A scheduled backup creates a local snapshot on the client's disk as a background task, without interrupting the end user's access to data and without moving the backup data across the network to a backup server.

The instant recovery process enables administrators to perform three different types of high-performance recovery operations:

- **Block-level restore:** Moves only blocks that have changed since the client's primary file set was backed up.
- **File promotion:** Restores a file from one of the disk volume snapshots to the original volume. Data that has undergone several changes since the last backup can be recovered more quickly using file promotion than block-level restore.
- **Image rollback:** Instantaneously restores the entire volume to a previous state and time.

Improving backup and recovery performance

Products such as VERITAS Backup Exec 9.1 for Windows Servers and VERITAS NetBackup 5.0 software can help administrators implement a flexible, efficient data protection and disaster recovery strategy. Disk-based backup and recovery operations help administrators improve performance substantially over traditional sequential tape-based backups while offloading the host CPU to help increase system availability for business-critical applications.

Scott Kosciuk (scott.kosciuk@veritas.com) is a product marketing manager for VERITAS NetBackup.

Michael Parker (michael.parker@veritas.com) is a product marketing manager for VERITAS Backup Exec.

Mark Thomason (mark.thomason@veritas.com) is a technical product manager for VERITAS Data Protection Solutions.

VERITAS Software Corporation (<http://www.veritas.com>) is a leading storage software company, providing data protection, application performance, storage management, high availability, and disaster recovery software.

FOR MORE INFORMATION

VERITAS Backup Exec for Windows Servers:

<http://www.veritas.com/products/category/ProductDetail.jhtml?productId=bews>

VERITAS NetBackup:

<http://www.veritas.com/products/category/ProductDetail.jhtml?productId=nbux>

VERITAS VISION Utility Computing Conference, May 3–7, 2004:

<http://www.veritas.com/vision>

Using disk-resident snapshots, tape backups, or both, administrators can reduce network backup windows and free host CPU and I/O cycles to process business-critical applications.

Leveraging the Microsoft Virtual Disk Service Using QLogic SANsurfer VDS Manager

The Microsoft® Windows Server™ 2003 operating system provides enhanced services for managing storage area networks—including Virtual Disk Service (VDS), a storage management application that enables administrators to manage volumes across heterogeneous disk arrays from a single point of control. This article explores how administrators can simplify storage management using VDS and the graphical user interface—based QLogic® SANsurfer® VDS Manager utility.

BY TIM LUSTIG AND KEITH HAGEMAN

In the scramble to provide adequate storage capacity for rapidly increasing volumes of data and applications, administrators often must configure disparate disk arrays from various vendors. Typically, they must use a separate storage management application for each vendor's disk array, incurring additional expenses for software tools and the staff training to use those management tools.

Storage management applications that provide a central point of control and consolidate disparate arrays in a storage area network (SAN) can help simplify IT administration and help lower total cost of ownership (TCO) for heterogeneous disk arrays across a network. Microsoft has introduced SAN services directly into the Microsoft® Windows Server™ 2003 operating system (OS), including the Microsoft Virtual Disk Service (VDS). By providing a common storage management specification, VDS enables third-party vendors to develop an application that allows administrators to manage all storage disk arrays from within the Windows OS, regardless of vendor. Administrators can create logical storage units (LUNs), RAID sets, and volumes to manage heterogeneous disk arrays, helping to simplify storage management and lower TCO.

Microsoft VDS provides administrators with remote session access to any server on the network and flexible

scripting to automate routines through a command-line interface (CLI). Further enhancing usability, the QLogic® SANsurfer® VDS Manager adds a wizard-based graphical user interface (GUI) to the VDS platform, which offers both expert and nonexpert administrators easy and intuitive device discovery, LUN configuration, health diagnostics for storage components, and RAID management. The QLogic GUI scales easily to manage disk and storage subsystems ranging from small SAN configurations to complex enterprise SANs, and can help streamline the storage management process.

Understanding VDS

The hardware-independent VDS storage engine, virtualized in the Windows Server 2003 OS, provides a uniform interface that enables administrators to manage block storage within disk arrays. Because VDS enables network administrators to manage hardware and software volumes within the OS, no specialized training is required; the look-and-feel of VDS is similar to that of server-based disk management features for local direct attach disks.

VDS architecture provides a layer of abstraction that connects storage hardware and SAN management tools (see Figure 1). From the common VDS interface,

administrators can detect, configure, and manage LUNs, configuring virtual disks for specific needs directly in the OS.

Support for diverse storage hardware

By providing a standard application programming interface (API) and implementing a uniform, open architecture for managing disks, VDS can act as a coordination service that enables support for a broad range of storage hardware. VDS enables administrators to discover and configure any storage device equipped with a VDS hardware provider—the software that makes a product VDS-compatible.

Along with the VDS service, Microsoft provides three tools to manage storage subsystems: two command-line utilities—DiskPart.exe and DiskRaid.exe—and a Microsoft Management Console (MMC) snap-in for Disk Management, which provides a common host environment for the third-party hardware provider. DiskPart.exe uses command sequences to manage storage system objects such as disks, partitions, and volumes; DiskRaid.exe uses command sequences to create, configure, and manage LUNs in the RAID storage subsystems (see Figure 2).

Administrators can take advantage of powerful scripting and remote-session functionality to manage disks and volumes using the DiskPart.exe and DiskRaid.exe tools. However, command-line utilities require administrators to memorize the commands and syntax—and administrators must possess expert domain knowledge of the SAN environment before they can exercise these tools effectively. Alternatively, the VDS architecture enables administrators to manage storage systems within Windows Server 2003 using third-party utilities such as QLogic SANsurfer VDS Manager, which offers a simple, efficient alternative to command-line coding.

Simplifying SAN management

QLogic SANsurfer VDS Manager is a point-and-click, GUI-based utility that allows administrators to discover supported storage devices and servers, including host bus adapters (HBAs), and permits configuration and monitoring of these resources from within Windows Server 2003. Features include discovery with a logical storage or server view, binding, volume and disk status, fault analysis, and RAID configuration and management. The QLogic GUI simplifies storage

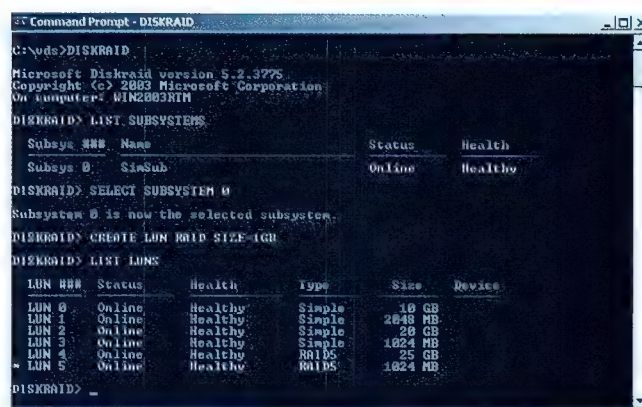


Figure 2. The DiskRaid.exe command-line utility

management further by providing wizards to create, configure, expand, shrink, and manage LUNs (see Figure 3). SANsurfer VDS Manager provides one simple interface for configuring cross-platform hardware, enabling autodiscovery of devices on the network using the APIs from QLogic HBAs. Microsoft-certified for Windows Server 2003, SANsurfer VDS Manager allows experts and nonexperts alike to access the full range of VDS features—simplifying storage management while helping to lower TCO.

Tim Lustig (tim.lustig@qlogic.com) is a business alliance manager for QLogic Corporation. He is the company's certification alliance program manager for the QLogic VDS project.

Keith Hageman (keithha@windows.microsoft.com) is program manager for the Enterprise Storage Division at Microsoft Corporation. He is the key technical spokesperson for VDS.

QLogic Corporation (<http://www.qlogic.com>) is a leading provider of controllers, host adapters, switches, and software for storage networks. QLogic has shipped more than 50 million products inside servers, workstations, RAID subsystems, tape libraries, and disk and tape drives.

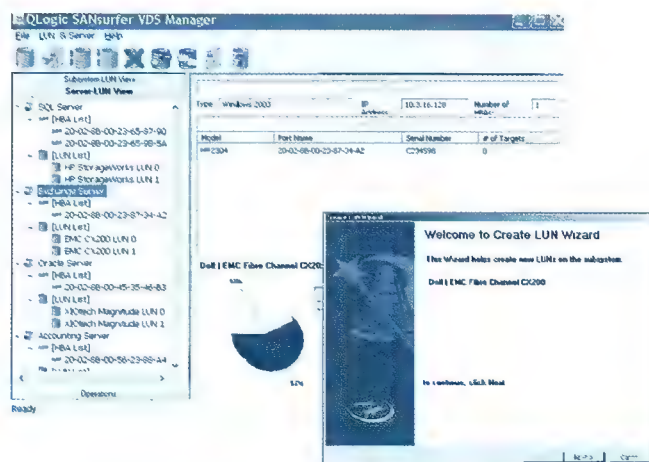


Figure 3. The wizard-based GUI for QLogic SANsurfer VDS Manager

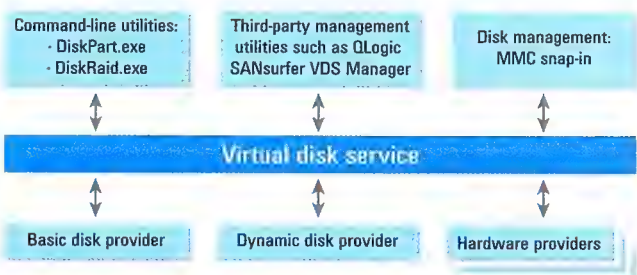


Figure 1. The three layers in VDS architecture

Simplifying Enterprise Backup and Restore Operations

Using BakBone NetVault

This article examines the deployment, configuration, and manageability of BakBone® NetVault™ tape backup and restore software. Three enterprise data backup scenarios are profiled, using NetVault with Dell™ PowerEdge™ servers, Dell PowerVault™ storage, and a Dell/EMC® storage area network.

BY JOE GALLO

BakBone® NetVault™ 7 software is based on a modular, object-oriented architecture that enables administrators to integrate tape backup and restore operations seamlessly with a variety of operating systems, databases and messaging applications, storage devices, and storage area networks (SANs). NetVault can be used to configure and deploy any storage device, as long as the client's operating system (OS) recognizes the device and can communicate with it through a network adapter. As enterprise IT environments grow and change, NetVault can help administrators scale out storage configurations quickly and flexibly to meet business needs.

A NetVault backup and restore configuration comprises three main components:

- **NetVault server:** A backup server that manages networked as well as direct attach storage devices, and also maintains a NetVault database of schedules, indexes, logs, and device information
- **Client:** A host server that participates in the networked backup and restore process
- **SmartClient™:** A NetVault client that also is configured with direct attach storage used in the backup and restore process

Using Dell™ PowerEdge™ servers, Dell PowerVault™ storage, and a Dell/EMC® SAN, BakBone engineers profiled three common enterprise computing configurations at Dell labs in October 2003. For each scenario, BakBone engineers

installed, configured, and demonstrated NetVault running a complete, error-free restore and backup operation.

Profile 1: Basic enterprise configuration

Once engineers completed the hardware setup for the basic enterprise configuration, installing the NetVault software took approximately 15 minutes. The NetVault software installation does not require administrators to reboot the servers, which is particularly advantageous in computing environments that cannot afford downtime.

Application Plugin Module™: To enhance NetVault support for third-party database and messaging applications, administrators can install an optional Application Plugin Module (APM™) from the NetVault server. BakBone offers APMs for Oracle®, Sybase®, Informix®, and various other applications. Using APMs, NetVault automatically adds application-specific components to the backup and restore selection criteria that appear on the NetVault graphical user interface (GUI). From the common NetVault GUI, administrators can manage all backup and restore operations across the SAN, network attached storage (NAS), wide area network (WAN), or local area network (LAN).

As enterprise computing environments become more complex, NetVault can help simplify administration through policy-based job management tools that allow IT staff to create editable, reusable job templates. Policy-based administration can improve productivity and help reduce human error by enabling administrators to monitor, manage,

and edit multiple jobs as a single group, or *set*—and apply job templates to additional hosts as environments grow.

Profile 2: Enterprise SAN configuration

A major objective in profiling the enterprise SAN environment was to determine the amount of NetVault reconfiguration that would be required to migrate from a direct attach SCSI Dell PowerVault 132T tape library to a PowerVault 132T connected to a SAN. Building upon the basic enterprise configuration in profile 1, BakBone engineers connected additional servers to the SAN, installed a PowerVault 132T Fibre Channel bridge module, and linked the bridge to a newly installed host bus adapter (HBA) in the NetVault server. For this migration, engineers performed three main NetVault configuration tasks: installing NetVault SmartClient software on the two new SAN servers; identifying the servers as clients within the NetVault server; and scanning for shared drives.

Engineers reused the policy management sets they created for the basic enterprise environment with slight modifications to the software configuration to recognize the new shared SAN storage. The NetVault process to reconfigure the direct attach storage model as the SAN storage model described for this scenario took approximately 30 minutes.

Many backup and restore offerings require administrators to install shared storage modules on every host that accesses SAN storage. These modules typically require licenses, which can increase storage costs. In comparison, NetVault software scans, discovers, and configures shared storage from the NetVault server instead of at each client—thereby helping to reduce installation time and lower costs. For the enterprise SAN configuration profile, engineers quickly configured NetVault to the shared storage option—which is required only on the NetVault backup server.

A modular, object-oriented approach to storage architecture enables NetVault to work with virtually all types of storage devices. Because many administrators are more familiar with backup and restore techniques involving tape drives and slots than physical disks,

NetVault displays physical disk space as virtual disk libraries (VDLs). Having a common interface for managing different types of storage media allows IT staff to implement new backup and restore strategies quickly and flexibly.

For the enterprise SAN configuration, BakBone engineers created a four-tape-drive, 30-slot VDL using a Dell/EMC CX600 Fibre Channel storage array. Engineers then created a backup policy to back up the VDL. The policy also specified that a copy from the VDL backup data be created on the PowerVault 132T tape library. The NetVault VDL software configuration and the creation of a backup policy were accomplished in minutes.

Profile 3: Enterprise NAS configuration

In an enterprise NAS environment, a NAS server such as the Dell PowerVault 775N can be used as the NetVault backup server (see Figure 1). After creating a direct SCSI connection between the PowerVault 132T tape library and the PowerVault 775N server, BakBone engineers installed the NetVault server software. The entire NAS configuration—including VDL and tape library initializations as well as policy management set creation—was completed in approximately three hours.

When a PowerVault 132T tape library is directly attached to a NAS device with NetVault installed, the maximum number of concurrent tape backup jobs is not restricted to the number of physical tape drives. That is, backup jobs can be copied to a VDL before—or instead of—the physical tape device. Because VDLs can be configured with an unlimited number of virtual tape drives, they overcome the limitations of a physical tape drive library. In this way, VDLs enable administrators to perform multiple backup jobs from different servers concurrently.

Implementing VDLs also can help administrators improve tape backup performance by eliminating the bottlenecks that can occur between data and tape when backup data streams are slow. After the initial backup, NetVault copies the backup data from the VDL to the physical tape library at SCSI bus speeds.

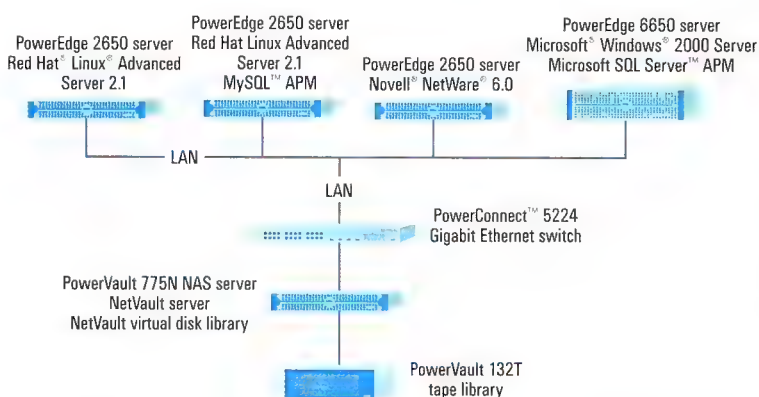



Figure 1. Enterprise configuration using NAS

Enterprise-class restore and backup

Modular, object-oriented architecture enables administrators to deploy various NetVault configurations quickly and easily. In addition, NetVault can help simplify systems management in heterogeneous environments by allowing IT staff to administer both tape-based and disk-based storage from a common GUI. Such flexibility and ease of use help make NetVault a cost-efficient backup and restore option for scalable enterprise computing environments. 

Joe Gallo (joseph.gallo@bakbone.com) is a senior sales engineer for BakBone Software, Inc. He specializes in developing backup and disaster recovery plans for distributed enterprise computing environments.

Migrating to Industry-Standard 64-bit Architectures

Deployments of industry-standard servers have grown quickly in the past decade. With the emergence of industry-standard 64-bit computing, many IT organizations no longer are deciding whether to migrate to a 64-bit platform, but when and where to migrate. This article can help administrators evaluate and begin planning that migration.

BY JOHN FRUEHE

The Intel® Xeon™ processor with 64-bit extension technology is designed to provide standards-based 64-bit computing and large-memory addressability with complete 32-bit compatibility. This innovation, along with other core technology enhancements, enables administrators to run 64-bit and 32-bit applications simultaneously while improving the performance of 32-bit applications. In addition, 64-bit extension technology helps IT organizations that purchase Intel Xeon processor-based servers to protect their capital investments. Administrators can migrate to 64-bit applications on the same hardware platform when software support becomes available.

As organizations review their IT strategy, they face a new decision-making process. For many, it is no longer whether they migrate to 64-bit servers built on an industry-standard platform, but where and when they will make this move. Migration planning has assumed strategic importance now that administrators have several 64-bit environments from which to choose. Previously, most IT organizations had not considered 64-bit computing critical for the countless infrastructure, Web, and general business applications because of unclear performance gains and the sheer scope of work involved in a comprehensive migration. Soon 64-bit extension technology from Intel

will help change the equation by bringing 64-bit computing to a new level of flexibility and price/performance.

Until now, cost and complexity have dictated how IT organizations deployed 32-bit and 64-bit systems: generally, infrastructure applications remained on 32-bit platforms while business logic applications ran on 64-bit platforms. Compared to proprietary RISC systems, a platform using industry-standard 64-bit Intel® Itanium® processors can offer compelling price/performance advantages for compute-intensive applications such as database transaction processing. As a result, many enterprises have conducted a cost/benefits analysis and decided to move business-critical applications to Itanium-based platforms.

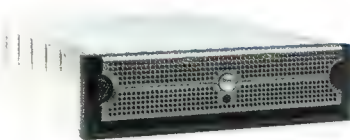
Operating systems, drivers, and applications—the complete platform software stack—must be modified to take advantage of 64-bit computing. For this reason, IT administrators can benefit from a practical, methodical migration approach. To map out a successful strategy for migrating to 64-bit computing, administrators should consider the following:

- What applications are good candidates for migration? Database and heavy-workload business applications usually top the list.



Dell/EMC STORAGE MAGNIFIED

MORE POWER _ MORE SCALABILITY _ MORE FLEXIBILITY



Dell/EMC CX300*

- 240% more sequential performance (MB/s)
- 100% more drives
- SnapView™ and Unix®** support

* Compared with CX200
** Solaris, AIX and HP-UX



Dell/EMC CX500*

- 100% more hosts
- 100% more disks
- 100% more transactional performance (IOPs)

* Compared with CX400



Dell/EMC CX700*

- 25% faster transactional processing (IOPs)
- Twice the number of backend ports
- 8GB of cache standard

* Compared with CX600

Uncover more information on Dell/EMC® products at www.dell.com/storage.

DELL | **EMC²**

Dell is not responsible for errors in typography or photography, or omissions. Dell and the Dell logo are trademarks of Dell Inc. EMC is a registered trademark and SnapView is a trademark of EMC Corporation in the United States of America. Other trademarks and trade names may be used in this document to refer to either the entities claiming the marks and names or their products. Dell disclaims proprietary interest in the marks and names of others. ©Copyright 2004 Dell Inc. All rights reserved. Reproduction in any manner whatsoever without the express written permission of Dell Inc. is strictly forbidden. For more information contact Dell. February 2004

- Which target architecture is best suited to the applications being migrated? Intel offers two industry-standard 64-bit architectures: the Intel Itanium 2 and Intel Xeon processors.
- Which operating system, applications, and drivers does the enterprise platform require—and are they all available in 64-bit versions? For example, the servers may run Microsoft® Windows® or Linux® operating systems; Microsoft SQL Server™, Oracle®, or SAP® databases; and RAID, Fibre Channel, or LAN storage environments.
- What services are required to assist in a successful transition?

The effect of processor architecture on applications

When choosing the most appropriate processor architecture, administrators first must determine the performance parameters of each application and how the application handles data. Memory size and data width of the processor execution path can be key performance factors for the overall system. However, their impact may vary from moderate to massive depending on the specific application and IT environment. Data-handling and memory size factors that help determine a suitable processor architecture include:

- **Sequential versus random requests:** Video decoding and streaming, for example, require a continuous set of sequential and structure calculations that take full advantage of 64-bit platform performance. In contrast, file-and-print sharing requires the processor to address multiple low-level requests from multiple users in a random fashion, making it less processor-dependent.
- **Logic-based versus load-based requests:** For example, life sciences applications tax the processor heavily by requiring large, complex algorithms and floating-point calculations. Domain Name System (DNS) and Secure Sockets Layer (SSL) applications use simpler algorithms but perform these calculations repeatedly.
- **Memory set:** The amount of memory that is utilized by the application can have a tremendous impact on overall system performance. For large memory requirements, 64-bit platforms help improve performance by providing a large addressable memory space.

Applications generally fall into three categories based on usage:

- **Compute-intensive:** Vertical and business-critical applications are included in this category, such as life sciences and high-performance computing; back-end database applications such as SQL Server, Oracle, and IBM DB2® software; business

applications such as customer relationship management (CRM) and enterprise resource planning (ERP); and e-business applications such as online commerce stores. These applications can benefit from 64-bit processors.

- **Compute/load-balanced:** Infrastructure-based applications are included in this category, such as Internet caching, security, DNS, Dynamic Host Configuration Protocol (DHCP), SSL, and database front ends. These applications may or may not benefit from 64-bit processors, so IT administrators should evaluate the environment before deciding to migrate.
- **Standard infrastructure:** Simple file-and-print sharing, resource sharing, and less critical single-use/low-volume business applications are included in this category. In general, standard infrastructure applications are less processor-intensive and thus will not benefit as much from 64-bit processing as compute-intensive and compute/load-balanced applications.

Once administrators have categorized applications, they can determine which of the three architectural platforms—32-bit, 64-bit extension technology, or 64-bit—best suits their needs. Figure 1 describes the three Intel processors that correspond to these platforms.

The IA-32 platform

Although a large portion of the existing software based on 32-bit Intel Architecture (IA-32) will not be migrated to 64 bits for some time, the 64-bit extension technology that will be included in Intel Xeon processors will maintain compatibility with this 32-bit software.¹ Several performance-enhancing innovations that Intel has added with 64-bit extension technology include faster CPU frequency, faster frontside bus speed, PCI Express™ I/O and graphics, and support for next-generation double data rate (DDR2) memory. However, many commercial and internally developed server applications were written on and for 32-bit architectures and have inherent limits therein, such as 2 GB memory space. Even as server consolidation occurs, such applications derive little benefit from the wider execution paths and increased memory capacity of 64-bit computing. Although they may eventually move to 64-bit extended architectures, these applications most likely will remain in 32-bit mode.

The 64-bit Intel Itanium 2 architecture

The Intel Itanium 2 architecture provides the top raw TPC-C performance of any Intel processor at a compelling price relative to RISC platforms—offering a robust, mature 64-bit environment for standards-based systems.² These characteristics make Itanium 2-based servers an appealing alternative to more expensive, proprietary 64-bit architectures.

¹ The 32-bit applications will need to be recertified to run on a 64-bit operating system.

² Results for both performance and price/performance are based on TPC-C benchmarks as of February 17, 2004. Current results can be found at <http://www.tpc.org>.

	Intel Xeon processor (32 bits)	Intel Xeon processor with 64-bit extension technology	Intel Itanium 2 processor (64 bits)
32-bit mode	Native	Native	Through emulation layer
64-bit mode	No	Through extension technology	Native
System bus	533 MHz, 64 bits wide; up to 4.3 GB/sec bandwidth	800 MHz, 64 bits wide; up to 4.3 GB/sec bandwidth	400 MHz, 128 bits wide; up to 6.4 GB/sec bandwidth
Cache (level 2/level 3)	512 KB/up to 2 MB	512 KB/up to 2 MB	512 KB/up to 6 MB
Memory addressing	32 bits (4 GB)	36–40 bits (1 TB)	50 bits (1024 TB, or 1 PB)
Error recovery on data bus (error-correcting code or retry)	No	No	Yes
Lockstep support	No	No	Yes
Corrupted data containment	No	No	Yes
Support for IBM Chipkill™ memory feature, retry on double bit	Yes	Yes	Yes
Memory spares	Yes	Yes	Yes

Figure 1. Comparing three Intel architectures: 32-bit Intel Xeon processor, Intel Xeon processor with 64-bit extension technology, and 64-bit Intel Itanium 2 processor

High-performance computing applications such as technical computations, life sciences, oil and gas research, and graphical rendering can take advantage of Itanium 2 architecture, mostly in dual-processor configurations. Business applications, databases, and other compute-intensive applications now are available to run on Itanium 2 versions of both Linux and Microsoft Windows Server™ 2003 operating systems.³ These applications have been developed and tuned specifically to run on Intel Itanium 2 processors and to take advantage of large 64-bit instructions and memory sets. Access to large amounts of memory helps these applications run faster because they use a flat memory address space and rely less on resource-intensive memory managers and paging to hard drives.

The Intel Xeon processor with 64-bit extension technology

By introducing 64-bit extension technology into the Intel Xeon processor family, Intel is creating a new class of 64-bit computing. This technology builds on the existing 32-bit Intel architecture and is expected to provide a more flexible, lower-cost platform than Itanium 2—supporting both 32-bit and 64-bit computing for applications that can benefit from features such as 64-bit operations and large addressable memory capacity.

The Intel Xeon architecture with 64-bit extension technology is designed to execute instructions at both 32-bit and 64-bit levels. As a result, it can be advantageous for mixed-purpose or infrastructure servers that are running a variety of applications on a single platform. Applications such as directory services, DNS, database front ends, and eventually messaging and groupware can benefit from 64-bit extension technology because it is expected to provide not only greater

performance than the existing 32-bit Intel Xeon architecture but, more importantly, better price/performance relative to Itanium 2 for the large number of servers often required to deliver such services.


Software support for 64-bit platforms

The key advantage of 64-bit extension technology is the ability to move applications from a 32-bit environment to a 64-bit environment on existing hardware as soon as software support becomes available. Therefore, administrators should assess the availability of software support before committing to a migration timeline.

Although legacy 32-bit operating systems, applications, and drivers are supported seamlessly on 64-bit extended platforms, to run in true 64-bit mode the entire software stack must be recompiled for 64-bit extension. Today, most Linux variants are available for 64-bit extended systems. Microsoft has announced that it plans to release a version of Windows Server 2003 for 64-bit extended systems to complement its 32-bit and Itanium versions of Windows Server 2003. *Note:* Itanium-based software does not run on 64-bit extended systems, and 64-bit extended software does not run on Itanium-based systems. Applications also must be written to support 64-bit extended systems.

Once administrators have determined the hardware platform, ensuring that appropriate drivers exist for all peripherals is a critical step in the migration process. Device drivers that were written for 32-bit operating systems will not work on 64-bit operating systems, including Itanium and Intel Xeon in 64-bit mode. Any 64-bit operating system, including Windows and Linux, will require its own set of drivers and many of those drivers will not be delivered by the operating system vendor, but rather by the device manufacturer.

The future of 64-bit technology

Intel offers two industry-standard 64-bit architectures for extending the IT infrastructure: Itanium 2, which provides high raw performance, and Intel Xeon with 64-bit extension technology, which provides excellent price/performance. At the same time, 32-bit applications will continue to have a place in the IT environment, regardless of whether they run on 32-bit or 64-bit extended systems. The flexibility in 64-bit extended systems enables 32-bit applications to run in a mixed environment alongside 64-bit applications on the same server. Because these different platforms require different operating systems, drivers, and applications, administrators must consider software support when planning a migration. From the hardware platforms to the drivers, operating systems, and applications, everything must be designed to work in a 64-bit environment to maximize performance. 

John Fruehe (john_fruehe@dell.com) is a marketing strategist for the Dell Enterprise Product Group. He has worked at Dell for more than 7 years; prior to that, he was at Compaq and Zenith Data Systems. John has a B.S. in Economics from Illinois State University and has been in the technology field for 13 years.

³ For a list of 64-bit applications certified to run on Intel Itanium 2 processors, visit <http://www.intel.com/cd/ids/developer/asmo-na/eng/catalog/processor/itanium/index.htm>.

Scaling Out Microsoft Exchange 2000 Server with Dell PowerEdge 6650 Servers

This article examines the scalability and performance of two Dell™ PowerEdge™ 6650 servers running Microsoft® Exchange 2000 Server. The four-processor Dell servers were connected to a Dell/EMC® CX600 storage array, which supported a storage area network. In tests using the Microsoft LoadSim 2000 tool, the PowerEdge 6650 servers sustained a maximum of 24,000 simulated Microsoft Outlook® users—indicating that the PowerEdge 6650 server can be an excellent platform for Exchange 2000 Server messaging environments.

BY FATIMA HUSSAIN AND SCOTT STANFORD

As messaging workloads increase in today's business environments, administrators typically implement one of two common approaches to satisfy the need for more processing and I/O capacity. *Scaling up* is based on a monolithic model that concentrates the workload on a single server containing eight or more processors. Although this approach can satisfy immediate and near-term processing needs, scaling up can require costly, proprietary systems that often do not have the logical or physical flexibility to meet rapidly changing business needs.

An alternative method, *scaling out*, follows a modular model that distributes the workload among a number of smaller servers configured with one, two, or four processors. This approach can help administrators create a dynamic and flexible architecture that comprises cost-effective, industry-standard servers to address constantly evolving business computing requirements. For messaging environments, scaling out can provide performance

advantages over scaling up—such as lower latency and higher messaging throughput.

This article explains how the modular approach can help administrators achieve high performance and scalability in an enterprise messaging environment. In July 2003, a Dell™ team tested a two-node configuration of Dell PowerEdge™ 6650 servers running Microsoft® Exchange 2000 Server. The default Microsoft LoadSim 2000 MAPI (Messaging Application Programming Interface) Messaging Benchmark 2 (MMB2) profile was used to simulate Microsoft Outlook® user actions and profiles representing various workload scenarios.

Configuring the test environment

The test environment consisted of two Dell PowerEdge 6650 servers, a Dell PowerEdge 1650 server, a Dell PowerConnect™ 5224 switch, a Dell/EMC® CX600 storage array, and 16 clients. All three servers used the Microsoft Windows® 2000 Advanced Server operating



More data? Less time? No problem.

Dell | Enterprise

Dell LTO-2 products keep pace with your growing backup needs.

Your organization's data may be increasing exponentially, but your backup window isn't. Chances are, your IT staff must perform backups overnight in only four to five hours—no matter how much data is involved.

Dell helps enterprises meet the challenges of rapid data growth and shrinking backup windows with the Dell™ PowerVault™ series of Ultrium® 2 LTO (LTO-2) tape backup products. Featuring powerful second-generation Linear Tape-Open™ (LTO®) technology, Dell LTO-2 products are optimized to provide both speed and capacity:

- **Outstanding performance:** A data transfer rate of up to 70 MB/sec.¹
- **High-capacity media:** Up to 400 GB capacity per cartridge can mean fewer cartridges to purchase and store, which can help save you money.²
- **Open standards:** LTO open standards foster competitive pricing and a format that has broad industry acceptance.

Best of all, the LTO roadmap projects the potential to double capacity and performance in each successive generation, and supports comprehensive backward compatibility.³ Dell LTO-2 drives are available in the PowerVault 110T (as a stand-alone unit or configured internally on select PowerEdge™ servers) and in Dell PowerVault 132T, 136T, and 160T LTO-2 tape libraries. To learn more, visit www.dell.com/storage.



Dell PowerVault 110T LTO-2 tape drive
(external model shown)



Dell PowerVault 132T, 136T, and 160T LTO-2 tape libraries



Click www.dell.com/storage

¹ Assumes 2:1 compression. Data compression rate may vary depending on settings, user environment and applications. For more information, visit http://www.dell.com/us/en/esg/topics/esg_tbu_main_storage_2_tapeb_110lto2.htm

² Assumes 2:1 compression. Data compression rate may vary depending on settings, user environment and applications. For more information about LTO format and capacity, visit <http://www.lto.org/news/ntm/format.htm>

³ For more information about LTO technology and the LTO roadmap, visit <http://www.lto.org> and <http://www.lto.org/news/ntm/roadmap.htm>

Linear Tape-Open (LTO) and Ultrium are trademarks or registered trademarks of Cerence, International Business Machines Corporation, and Hewlett-Packard Company. Dell, the Dell logo, PowerEdge, and PowerVault are registered trademarks of Dell Inc. ©2004 Dell Inc. All rights reserved.

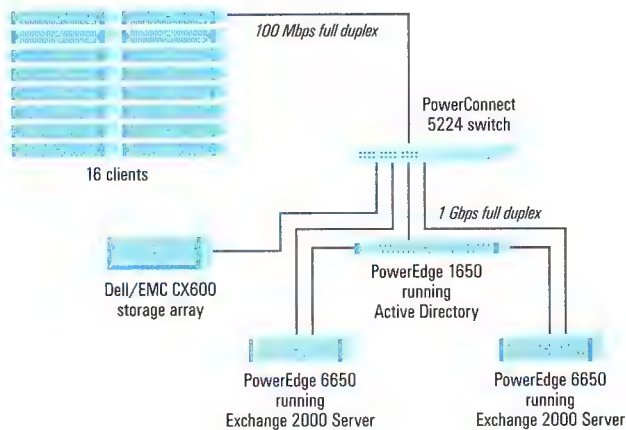


Figure 1. Network configuration for Microsoft Exchange 2000 messaging environment

TUNING WINDOWS 2000 ADVANCED SERVER

To tune the Windows 2000 Advanced Server OS for Exchange 2000 Server, the Dell test team added the `/3GB` switch to `boot.ini`:

```
[boot loader]
timeout=30
default=multi(0)disk(0)rdisk(0)partition(2)\WINNT
[operating systems]
multi(0)disk(0)rdisk(0)partition(2)\WINNT="Microsoft
Windows 2000 Advanced Server" /3GB /fastdetect
```

For more information, see the Microsoft Knowledge Base article (ID number 266096) online at <http://support.microsoft.com>.

The team allocated the following sizes to NTFS units:

- Information Store transaction logs: 8 KB
- Information Store databases: 16 KB
- Operating system: default

The `HeapDecommitFreeBlockThreshold` registry key enables Exchange administrators to better control how memory is handled as it frees up. To help mitigate the potential for virtual address fragmentation, the Dell test team set this key to the Microsoft-recommended value: under `HKEY_LOCAL_MACHINE\SYSTEM\CurrentControlSet\Control\Session Manager`, the team set the `HeapDecommitFreeBlockThreshold` data type to `REG_DWORD` and the value to `0x00040000` (hex).

For more information, see the Microsoft Knowledge Base article (ID number 315407) online at <http://support.microsoft.com>.

system (OS). The two PowerEdge 6650 servers ran Microsoft Exchange 2000 Server, Enterprise Edition; the PowerEdge 1650 server ran the Microsoft Active Directory® directory service.

Each PowerEdge 6650 server was configured with four Intel® Xeon™ processors MP at 2.0 GHz and 2 MB level 3 (L3) cache. Other configuration settings for the PowerEdge 6650 were as follows:

- 4 GB double data rate (DDR) SDRAM
- Eight 512 MB dual in-line memory modules (DIMMs)
- One Intel PRO/1000 XT Gigabit Ethernet¹ network interface card (NIC)
- Two QLogic® 2340 host bus adapters (HBAs)
- One PowerEdge Expandable RAID Controller, Dual Channel (PERC 3/DC) with 128 MB cache and internal and external channels
- Two 18 GB Ultra320 SCSI hard drives at 15,000 rpm

The PowerEdge 6650 servers supported two storage groups, four mail databases, and two transaction logs. Public folder information was housed in one of the database logical storage units (LUNs) on each server. The PowerEdge 6650 servers used hardware RAID-1 containers for the OS and Exchange files.

The 1U PowerEdge 1650 was configured with 4 GB of RAM, an embedded Gigabit Ethernet network adapter, and two Intel® Pentium® III processors at 1.4 GHz with 256 MB of level 2 (L2) cache. Windows 2000 Advanced Server and the Active Directory database and logs were housed on a RAID-1 mirrored volume controlled by the server's embedded PERC 3, Dual Channel integrated (PERC 3/Di). The third disk in the PowerEdge 1650 was dedicated as a hot spare.

As shown in Figure 1, the servers and clients were interconnected using a Dell PowerConnect 5224 Gigabit Ethernet switch. The Active Directory and Exchange servers were linked at 1 Gbps full duplex to minimize any network-related latency. All clients were connected at 100 Mbps full duplex. The test team used Dell OpenManage™ Server Assistant version 7.4 to configure all three servers, and Dell OpenManage Systems Management version 3.4 to administer them.

Minimizing bottlenecks in the test configuration

To focus on measuring performance and scalability at the processor level, the test team configured the storage subsystem to reduce I/O-related bottlenecks. Although RAID-10 and RAID-5 are common fault-tolerant options for databases and transaction logs, the additional latency inherent in fault-tolerant RAID subsystems can create a performance bottleneck under heavier

¹ This term does not connote an actual operating speed of 1 Gbps. For high-speed transmission, connection to a Gigabit Ethernet server and network infrastructure is required.

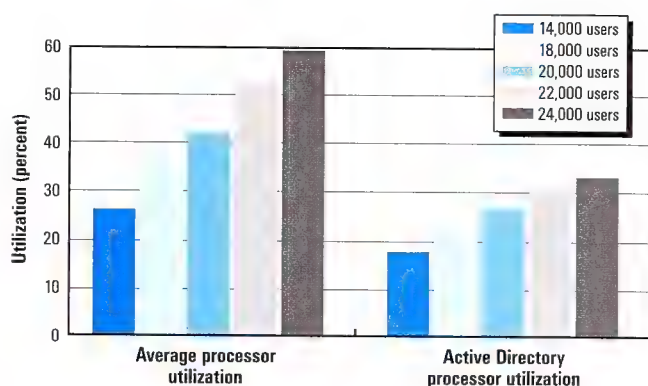


Figure 2. Processor utilization for two-node PowerEdge 6650 configuration as Exchange 2000 Server workload increased

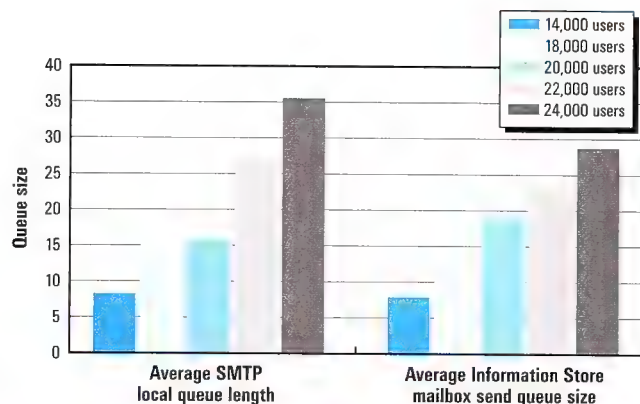


Figure 3. SMTP and Information Store queues for two-node PowerEdge 6650 configuration as Exchange 2000 Server workload increased

workloads. For that reason, the Dell team tested the scale-out configuration using RAID-0 for the database and transaction log LUNs, which shifted any potential resource bottlenecks to the processors.

The team also tuned Microsoft Windows 2000 Advanced Server, Microsoft Active Directory, Microsoft Exchange 2000 Server, and the network adapter settings to minimize any other potential bottlenecks. For more information, see "Tuning Windows 2000 Advanced Server" and "Optimizing Exchange 2000 Server."

Building the storage area network

To build the storage area network (SAN), the test team used a Dell/EMC CX600 storage array, which can support up to 16 Fibre Channel or ATA disk array enclosures. The Dell team used Dell/EMC DAE2 disk array enclosures, which can hold up to 15 drives.

Using EMC Access Logix™ software to provide LUN masking between the two PowerEdge 6650 servers, the test team created two storage groups, one for each server. Within each storage group, one Dell/EMC Fibre Channel DAE2 disk array enclosure hosted a 12-drive LUN. The team used this LUN to create two dynamic NT file system (NTFS) volumes to support each server's Exchange transaction logs. Eight additional DAE2 enclosures were allocated evenly between the two storage groups so that each server had 60 drives available for four 15-drive database LUNs. From these four LUNs per host, four dynamic NTFS volumes were created to support each of the mail databases. For more

For messaging environments, scaling out can provide performance advantages over scaling up—such as lower latency and higher messaging throughput.

details on the SAN configuration, see "Configuring the Dell/EMC CX600 SAN."

Analyzing the performance data

The Dell test team focused on the scalability and performance of the two PowerEdge 6650 servers as the Exchange workload increased. Exchange 2000 Server data and OS performance monitor (perfmon) data were used to evaluate system performance using critical application and server health parameters, and to determine at what point messaging workloads were no longer sustainable. Processor utilization, the Exchange Information Store mailbox send queue size, Simple Mail Transport Protocol (SMTP) local queue lengths, and 95th percentile response times² were some of the key metrics used for this comparative analysis.

Figure 2 presents processor utilization across the two-node configuration as workload increased, showing both the average utilization and the utilization for the Active Directory service. Figure 3 illustrates the effect of increasing workload on SMTP and the Information Store, and Figure 4 shows the effect that scaling out has on

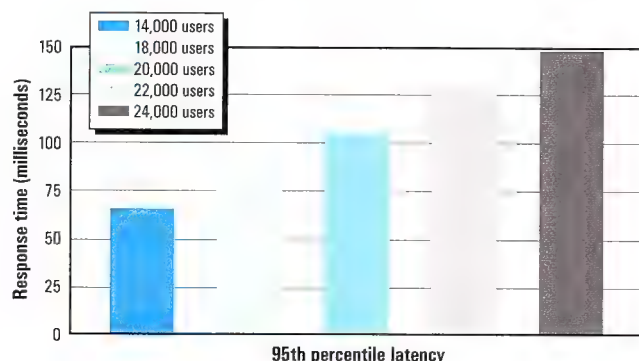


Figure 4. Latency (95th percentile response times) for two-node PowerEdge 6650 configuration as Exchange 2000 Server workload increased

² The 95th percentile response time is a measurement used to determine how quickly the Exchange 2000 Server application responds to LoadSim 2000 client actions. Response time is measured in milliseconds.

latency. The two-node PowerEdge 6650 configuration performed well as the workload increased from 14,000 to 24,000 MMB2 users, with slight performance degradation under heavier loads.

For the test results presented in Figures 2, 3, and 4, the Dell team enabled Intel Hyper-Threading Technology—a feature of the Intel Xeon processor. Hyper-Threading Technology can help improve processor resource utilization by making one physical processor appear to the OS as two processors.³ To further illustrate the performance capabilities of the PowerEdge 6650 server, the test team compared the performance of the two-node configuration with Hyper-Threading enabled and disabled. As the results in Figure 5 show, Exchange 2000 Server performance was significantly enhanced when Hyper-Threading Technology was enabled.⁴

Scaling out Exchange 2000 to attain key advantages

IT organizations can benefit in several ways from scaling out Exchange 2000 Server on Dell PowerEdge 6650 servers running Windows 2000 Advanced Server:

- **Extensible, modular design:** The two-node PowerEdge 6650 configuration can support various Active Directory and IP site

IT organizations can benefit from extensible, modular design when scaling out Exchange 2000 Server on Dell PowerEdge 6650 servers running Windows 2000 Advanced Server.

	12,000 MMB2 users	14,000 MMB2 users	14,000 MMB2 users
Hyper-Threading enabled	No	No	Yes
Average processor utilization (%)	35.20	41.06	26.28
Average SMTP local queue length	6.10	9.41	8.32
Average Information Store mailbox send queue size	5.60	9.20	7.96
95% latency (milliseconds)	81	80	78
Active Directory processor utilization (%)	15.40	17.93	17.74

Figure 5. Exchange 2000 Server performance for two-node PowerEdge 6650 configuration with Hyper-Threading Technology enabled and disabled

topologies, connectors, and Exchange 2000 organizational and routing group boundaries.

- **Cost-effective memory architecture:** Distributed nodes with two to four processors and 4 GB of RAM more closely match the Exchange 2000 memory architecture. Systems with more than four processors typically need greater memory to support those processors even though Exchange 2000 uses only 2.8 GB of RAM per server.
- **Fault tolerance:** Setting up two nodes as a failover cluster can help prevent data loss. If one server fails, the other node can take responsibility for the failed server.
- **Intel Hyper-Threading Technology:** When running Windows 2000 Advanced Server, the four-processor PowerEdge 6650 server can take advantage of the Hyper-Threading Technology of its Intel Xeon processors, so that the four physical processors act as eight virtual processors. Servers with eight physical processors must

OPTIMIZING EXCHANGE 2000 SERVER

To further tune Exchange 2000 Server, the Dell test team set the msExchESEParamLogBuffers integer value to 500. Tuning can be performed regardless of configuration—whether in a single-node, scale-out, or clustered environment. For more information, see the *Microsoft Exchange 2000 Internals: Quick Tuning Guide* online at <http://www.microsoft.com/technet/treeview/default.asp?url=/technet/prodtechnol/exchange/exchange2000/maintain/optimize/exchtune.asp>.

To optimize Exchange 2000 Server for the Windows 2000 Server OS, the team configured the DSAccessCache by adjusting the cache expiration time, the maximum cache size (memory), and the maximum number

of entries. Under HKEY_LOCAL_MACHINE\System\CurrentControlSet\Service\MSExchangeDSAccess\Instance0, the team set the following data types and values:

- CacheTTL: REG_DWORD, 0x600
- MaxMemory: REG_DWORD, 0x3200000
- MaxEntries: REG_DWORD, 0x0

For more information, see the *Microsoft Exchange 2000 Server Resource Kit*, pp. 901–903.

³ For more information about the potential benefits of Hyper-Threading Technology, please visit <http://www.intel.com/ebusiness/hyperthreading/server/index.htm>.

⁴ For more information about the potential benefits of Hyper-Threading Technology for Exchange workloads, see "Impact of Hyper-Threading Technology on Exchange 2000 Performance" by Scott Stanford and Ramesh Radhakrishnan, Ph.D., in *Dell Power Solutions*, August 2002.

Maximum ROI. Minimum IOU.



Dell | Enterprise

The power of Dell flexibility.

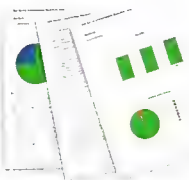
What does Dell bring to your enterprise? Just what you'd expect: A legendary focus on you, the customer, that's as relentless as our focus on driving down costs.

With Dell you get:

- **Enterprise solutions on your terms.** No proprietary systems, no endless consulting fees. Just powerful, cost-effective, industry-standard technology like Dell™ PowerEdge™ servers with Intel® Xeon™ processors.
- **Simplified systems management.** From SAN and server consolidation to UNIX® migration, we collate flexible systems that are easier to afford, year after year.
- **Enterprise services.** Dell has an expanded range of services to help you simplify the design and management of your IT infrastructure.



Dell PowerEdge servers use Intel® Xeon™ processors.



For nearly 20 years, we've revolutionized the way the world buys and manages technology. Now find out how Dell's direct approach can revolutionize your enterprise. To learn more about the Dell ROI test, visit www.dell.com/ROI1.



Enterprise solutions that can cut costs today and tomorrow. Easy as



Visit www.dell.com/ROI1

CONFIGURING THE DELL/EMC CX600 SAN

Several configuration settings can help optimize SAN performance within an Exchange 2000 Server messaging environment. The Dell test team configured the cache for the Dell/EMC CX600 storage array with the following settings:

- Page size: 4 KB
- Low watermark: 60
- High watermark: 80
- Storage processor: enabled for read and write caches
- Logical disks: write cache enabled; read cache disabled

Memory for the SAN was allocated as follows:


- Storage processor memory: 8 GB total
- Read cache memory: 100 MB per storage processor
- Write cache memory: 2 GB per storage processor

The Dell/EMC CX600 storage array ran the following software revisions:

- Base software: 02.04.160.5.007
- Management user interface: 6.4.1.0.0
- Management server 6.4.0.5.2
- EMC Navisphere® Manager: 6.4.0.5.2
- Navisphere agent and command-line interface (CLI): 6.4.0 (5.2)

use the Windows 2000 Datacenter Server OS to take advantage of Hyper-Threading Technology, because Windows 2000 Advanced Server cannot support the 16 virtual processors that would be created.

- **Reduced Active Directory utilization:** Because each server can have its own cache, localized DSAccessCache instances on each Exchange server help to reduce Active Directory utilization and improve response times for client requests. In a monolithic server configuration, all clients are dependent on a single DSAccessCache.

The testing performed by the Dell team demonstrates how a modular approach can help administrators improve the performance and scalability of a Microsoft Exchange 2000 messaging environment. By scaling out PowerEdge 6650 servers, administrators can achieve lower latency and higher messaging throughput while building a flexible IT infrastructure that is responsive to ever-changing computing needs. Scaling out a cluster with one-, two-, or four-processor servers can be more cost-effective by using industry-standard components; more flexible by allowing servers to be added or removed easily; and more fault-tolerant by avoiding the single point of failure inherent in a scale-up configuration. 

Fatima Hussain (fatima_hussain@dell.com) is a systems engineer and a senior analyst in the Server and Storage Performance Analysis Group at Dell. She has worked on Exchange benchmarking and scale-out studies, but her current work focuses on running microbenchmarks—such as Linpack, Stream, SPEC^{int} CPU2000, SPECjbb^{int}, and LMBench—on current and next-generation server platforms to identify bottlenecks in the system architecture and to compare performance across systems. Fatima has a B.S. in Electrical and Computer Engineering from The University of Texas at Austin.

Scott Stanford (scott_stanford@dell.com) is a systems engineer in the Server and Storage Performance Analysis Group at Dell. His current work focuses on Exchange Server benchmarking. Scott served in the U.S. Peace Corps in Nepal and with the U.S. Army, 24th and 3rd Infantry Divisions. Before Dell, he worked in the public sector as an information services manager. He has an M.S. in Community and Regional Planning from The University of Texas at Austin and a B.S. from Texas A&M University. Scott is A+ and N+ certified and a Microsoft Certified Systems Engineer (MCSE).

FOR MORE INFORMATION

Dell PowerEdge 6650:

http://www1.us.dell.com/content/products/productdetails.aspx/pedge_6650?c=us&cs=555&l=en&s=biz

Microsoft Exchange Server:

<http://www.microsoft.com/exchange>

Explore
Dell Power Solutions

[HTTP://WWW.DELL.COM/POWERSOLUTIONS](http://www.dell.com/powersolutions)

Scaling Out Web Server Performance

on Dell PowerEdge 6650 Servers

A Dell™ team tested Web server performance on Dell PowerEdge™ servers using the industry-standard SPECweb® 99_SSL Web server benchmark. Measuring performance of a single PowerEdge 6650 server and then comparing it to two clustered PowerEdge 6650 servers indicated that near-linear scaling can be achieved.

BY DAVID J. MORSE

As organizations grow, they may need to scale out their Web sites by adding more servers. To test the Web-serving performance of Dell™ PowerEdge™ 6650 servers, in September 2003 the Dell Server Performance and Analysis team used an industry-standard benchmark to simulate heavy Web server workloads, first on a single server and then on a two-node cluster. The benchmark results showed that a single PowerEdge 6650 could sustain 2,177 simultaneous Secure Sockets Layer (SSL) Web server connections. A cluster of two identically configured PowerEdge 6650 servers sustained 4,224 connections—a scaling factor of 1.94. This near-linear scaling indicates that administrators can achieve excellent price/performance by adding servers to their environments as workload demand grows.

Exploring benchmark results for a single server

To test Dell PowerEdge 6650 performance, the Dell team chose the SPECweb® 99_SSL benchmark, an industry-standard benchmark developed by a consortium of software and hardware vendors. The SPECweb99_SSL benchmark

stresses several subsystems—especially CPU and networking—by simulating a typical Web server workload and measuring the maximum number of simultaneous connections that a server can sustain. The SPECweb99_SSL benchmark, which is based on the SPECweb99 benchmark, adds SSL protocol support.

For the first test, the team used one Dell PowerEdge 6650 server with four Intel® Xeon™ processors MP at 2.8 GHz and 16 GB of RAM, running the Red Hat® Linux® 8.0 operating system. The team selected Zeus Web Server™ 4.2 r2 for its ability to sustain high HTTP SSL loads.

Because SPECweb99_SSL primarily stresses the CPU and network subsystems, the internal storage in the PowerEdge 6650 was sufficient. The team used one 36 GB, 15,000 rpm drive for the operating system and configured four 18 GB, 15,000 rpm disks as a software RAID-0 ext2 file system for the Web server file set and logs.

The team used one Dell PowerConnect™ 5224 network switch. One Intel PRO/1000 MT Dual Port Server Adapter received requests from seven clients, which simulated Web browsers to put the server under a heavy load. This

configuration enabled the single-server PowerEdge 6650 to sustain 2,177 simultaneous SPECweb99_SSL connections.¹

Exploring benchmark results for two servers

For the test of the two-node cluster, the team used the same server hardware and software: two PowerEdge 6650 servers, each with four Intel Xeon processors MP at 2.8 GHz and 16 GB of RAM, running Red Hat Linux 8.0 and Zeus Web Server 4.2 r2.

One important addition, however, was a dedicated Gigabit Ethernet² link between the two Web servers. This link allowed files that changed dynamically during the run to be stored locally on one server but remain accessible to the other through Network File System (NFS), a network file system commonly used in UNIX® environments. The Dell team configured half of the clients to request Web pages from the first Web server and configured the other half to request pages from the second server. This arrangement resembles a real-world scenario in which a network load balancer, such as a PowerEdge Load Balancing Server-BIG-IP® Powered, routes browser requests intelligently to the least utilized server. Figure 1 shows the layout of the two-node cluster.

Using two PowerEdge 6650 servers, the cluster sustained 4,224 simultaneous connections.³ As shown in Figure 2, near-linear scaling was achieved by adding the second Web server—a scaling factor of 1.94 over the single-server results.

During the implementation of the two-node cluster, the Dell team addressed several hurdles relating to static and dynamic content of the Web server files.

Static content. Most of the Web server file set for SPECweb99_SSL consists of HTML pages that clients request, and 70 percent of the workload consists of static requests. Numerous directories exist, and clients request more files from more directories as the number of connections requested from the Web server increases.

The Dell team used the Linux rsync utility to synchronize the file sets, because each server needed to store the same static files locally. Using rsync ensured that client 1 requesting data from the first PowerEdge 6650 would receive the same response as client 2 requesting data from the second PowerEdge 6650.

Both servers were equipped with 16 GB of RAM to assist in

The SPECweb99_SSL benchmark stresses several subsystems—especially CPU and networking—by simulating a typical Web server workload.

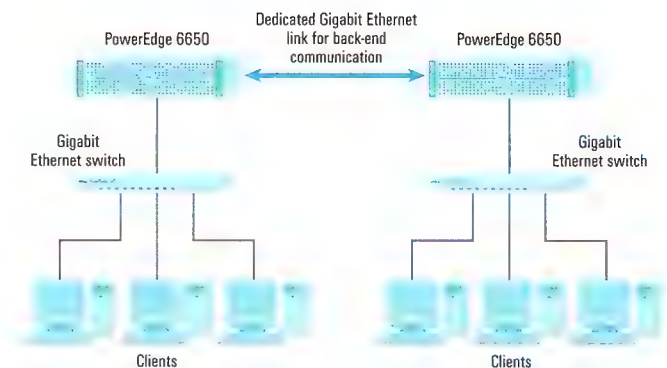


Figure 1. Web server benchmark configuration for two-node cluster

caching the static files, because pages can be served much more quickly from RAM than from the disk subsystem. The file set was stored on a four-disk software RAID-0 stripe to ensure the fastest disk access times.

Dynamic content. The other 30 percent of the SPECweb99_SSL workload consists of dynamic requests; these requests stress the Web server by requiring it to execute Internet Server Application Programming Interface (ISAPI) server extensions and Common Gateway Interface (CGI) executables. The benchmark's dynamic content simulates two features common to commercial Web servers: advertising and user registration. Cookies are passed to and from the client and server, and the client also simulates form submission through a certain percentage of HTTP POST operations. The Web server records all user-submitted form data, keeps track of which advertisements users have seen, and rotates them so that it does not display the same advertisement twice. Some dynamic requests also require scanning the static HTML files and inserting certain strings; this process simulates server-side includes (SSIs), a common practice for real-world Web servers.

The SPECweb99_SSL dynamic content presented a twofold challenge: some files changed on the fly during a run, and the benchmark has a single point of contention—a POST log file—to prevent shared-nothing configurations. These circumstances present a problem for clustered configurations, because any changes to files on one Web server must be accounted for on other nodes so that the browser receives a unified, consistent view.

To solve the first problem, the second Web server accessed the dynamic files from the first server by using a network-mounted file system. The second problem, synchronizing writes from two servers to one POST log file, was more difficult to solve. Every time a client submits form data through the POST operation, the server

¹ These results have been submitted to SPEC and are available at <http://www.spec.org/osg/web99ssl/results/res2003q3>.

² This term does not connote an actual operating speed of 1 Gbps. For high-speed transmission, connection to a Gigabit Ethernet server and network infrastructure is required.

³ These results have been submitted to SPEC and are available at <http://www.spec.org/osg/web99ssl/results/res2003q3>.

must open one POST log file exclusively, write a line of data, and then close the log file. The benchmark rules state that the data must be logged in the order in which it was received. The Dell team developed code to allow the second server to send the necessary data to the first server through the back-end Gigabit Ethernet link. Doing so allowed for optimal performance and helped ensure that the entries were logged into one file in the proper order.

Achieving excellent scaling in high-stress environments

The single-server SPECweb99_SSL result of 2,177 connections shows that the PowerEdge 6650 can perform well under heavy-stress workloads. The two-server result is even more compelling, because it indicates that excellent performance can be obtained by clustering multiple PowerEdge servers as demand grows. This capability is especially important for the customer-facing Web server tier, which can quickly become a bottleneck as site traffic increases.

Dell products can help administrators manage end-user demands easily and effectively by scaling out enterprise environments to suit business needs. As the results in this article show, workloads can be balanced across multiple systems to improve performance. Multiple systems also help ensure availability if one server in the cluster goes offline because of scheduled hardware or software upgrades, system rebuilding, backups, or unexpected failures. Dell PowerEdge systems are an efficient choice for scale-out servers because they incorporate industry-standard software and hardware, cost-effectively providing both reliability and high performance. ☺

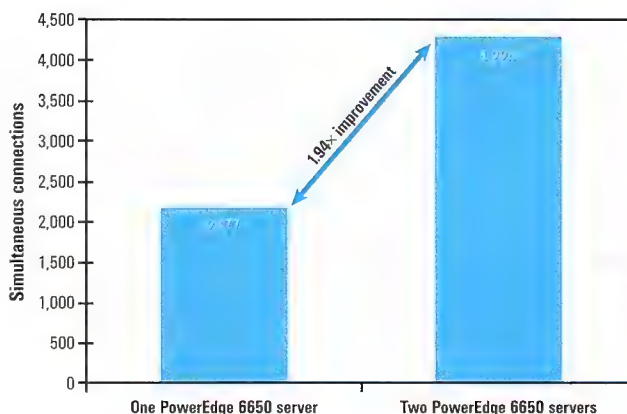


Figure 2. SPECweb99_SSL results for PowerEdge 6650 servers: one server versus two servers

David J. Morse (david_j_morse@dell.com) is a senior performance engineer for the Dell Server Performance and Analysis Lab. He specializes in Web server performance and is responsible for running the industry-standard SPECweb Web server benchmark across the Dell PowerEdge server product line. Before joining Dell, he spent two years at NCR in the performance integration and testing group. David has a B.S. in Computer Engineering from the University of South Carolina and is a Red Hat Certified Engineer (RHCE).

FOR MORE INFORMATION

SPECweb99_SSL benchmark:
<http://www.spec.org/web99ssl>

Share Your
Experience in
Dell Power Solutions

Dell Power Solutions is a peer-to-peer communication forum. We welcome subject-matter experts, end users, business partners, Dell engineers, and customers to share best-practices information. Our goal is to build a repository of solution white papers to improve the quality of IT.

Guidelines for submitting articles to *Dell Power Solutions* can be found at <http://www.dell.com/powersolutions>.

Scalable Enterprise Computing:

Testing a Clustered Database on the Dell PowerEdge 6650

Clusters of small, industry-standard servers offer several advantages over large, proprietary stand-alone servers. Such clusters can provide redundancy for high availability, lower hardware costs for the same number of processors, and the ability to increase processing power simply by adding another inexpensive server to the cluster. This article compares the performance of a cluster of Dell™ PowerEdge™ servers with that of a stand-alone IBM® server.

BY DAVE JAFFE, PH.D., AND TODD MUIRHEAD

Until recently, data centers required expensive, proprietary servers using eight or more processors to handle the intense computing demands of enterprise applications. However, advances in computer clustering technology and high-speed network interconnects now enable clusters of smaller servers using four or fewer processors to handle large-scale applications such as enterprise databases.

System administrators can cost-effectively increase the processing power available to enterprise applications by adding more servers, or nodes, to *scale out* a cluster. In comparison, to *scale up* processing capacity, administrators must either increase the number of processors in a single server or replace that server entirely with a higher-performance system. A cluster of smaller, industry-standard servers can be more cost-effective than a single, proprietary server—and a cluster of redundant servers can provide higher availability than a larger, stand-alone server because the cluster can be configured with no single point of failure.

Scaling out versus scaling up

To demonstrate the benefits of scaling out versus scaling up, in November 2003 a team of Dell engineers tested a

two-node cluster comprising two Dell™ PowerEdge™ 6650 servers, each with four Intel® Xeon™ processors MP running at 2.8 GHz and with 2 MB level 3 (L3) cache. The team compared the results to tests of an IBM® eServer® xSeries® 445 server with eight Intel Xeon processors MP at 2.8 GHz and 2 MB L3 cache. Both nodes of the Dell cluster and the IBM server ran a leading enterprise database server on the Microsoft® Windows® 2000 Advanced Server operating system.

In addition, the Dell configuration ran clustering software that enabled the two nodes to share a single database instance and ensured data coherency between the nodes. The Dell cluster and the eight-processor IBM system were connected to a Dell/EMC® storage area network (SAN) with several hundred gigabytes of available storage.

The test team used the Dell/EMC SAN to build identical, large (approximately 100 GB) databases representing an online DVD store. The team wrote software to simulate both online order entry and report generation, including a calculation of DVD sales by category for the previous month, quarter, and half year. The order-entry and report-generation workloads ran simultaneously in a typical enterprise customer scenario. Server CPU utilization

was kept below 90 percent to allow for order spikes, which also is typical of enterprise server use.

Putting servers to the test

To make a fair comparison, engineers configured the Dell cluster and the IBM system as similarly as possible—that is, the Dell cluster as a whole contained eight CPUs and 8 GB of memory, as did the stand-alone IBM system. In addition, each node in the Dell cluster required an additional Gigabit Ethernet¹ network interface card (NIC). The two nodes were connected by a Dell PowerConnect™ 5224 Gigabit Ethernet switch. Figure 1 shows the configuration for each server in the test configuration.

Each server was outfitted with two QLogic® 2340 host bus adapters (HBAs) for connection to the Dell/EMC SAN. The test team used EMC PowerPath® software to combine the two HBAs in each server into a fault-tolerant, load-balanced pair. The fact that the Dell cluster had twice as many connections to the back-end storage as the IBM server was immaterial to the performance measured in the test because none of the HBAs were heavily loaded.

The test team used the Dell/EMC SAN to store database tables for both the Dell cluster and the IBM server. Identical logical storage units (LUNs) were defined for both the cluster and the server on the SAN. The SAN components and disk layout are shown in Figure 2.

Setting up the enterprise database server

A leading enterprise database server using clustering technology enabled the two Dell servers to scale out as a cluster; the eight-processor IBM server was configured as a single-node database server.

Windows 2000 can address only 4 GB of memory. On the Dell and IBM servers, which were running Windows 2000 Advanced Server, the test team increased the maximum amount of memory the database could address by including the /3GB parameter in the boot.ini file. By specifying the /3GB parameter, administrators can enable 3 GB of memory to be used for applications and the remaining 1 GB to be used for the operating system. In addition, the test team set parameters enabling the database to support Address Windowing Extensions (AWE), which allow applications to address more than 4 GB of memory. These parameters enabled the IBM system to address 8 GB of memory.

On the Dell cluster and the IBM server—both Intel processor-based systems—the test team installed version 8.2.2.25 of the QLogic 2340 driver for Windows

A cluster of smaller,
industry-standard
servers can be more
cost-effective than a
single, proprietary server.

	Each Dell PowerEdge 6650	IBM eServer xSeries 445
Operating system	Microsoft Windows 2000 Advanced Server	Microsoft Windows 2000 Advanced Server
CPU	Four Intel Xeon processors MP at 2.8 GHz with 2 MB L3 cache	Eight Intel Xeon processors MP at 2.8 GHz with 2 MB L3 cache
Memory	4 GB	8 GB
Internal disks	Two 18 GB drives	Two 18 GB drives
NICs	Two 10/100/1000 Mbps (internal) NICs and one Intel PRO/1000 XT Server Adapter	Two 10/100/1000 Mbps (internal) NICs
Disk controller	PowerEdge Expandable RAID Controller 3/Dual Channel (PERC 3/DC)	IBM ServeRAID™ controller
Fibre Channel HBAs	Two QLogic 2340 HBAs	Two QLogic 2340 HBAs
Remote management card	Dell Remote Access Card III (DRAC III)	IBM Remote Supervisor Adapter
Video	On-board	On-board
Height	4U (7 inches)	4U (7 inches)
Cluster interconnect	PowerConnect 5224 Gigabit Ethernet switch	Not required

Figure 1. Configuration for Dell PowerEdge 6650 cluster and IBM eServer xSeries 445 server

Controller	One Dell/EMC CX600 storage array
Disk enclosures	Two Dell/EMC DAE2 disk array enclosures
Disks	Thirty 73 GB drives at 10,000 rpm
LUNs	Two 10-disk RAID-10 LUNs for data Two 2-disk RAID-1 LUNs for logs One 5-disk RAID-0 LUN for temporary data staging when loading One hot-spare disk
Software	EMC Navisphere® Manager EMC Access Logix™ EMC PowerPath

Figure 2. Configuration for Dell/EMC SAN in test environment

and PowerPath 3.0.5 for Windows. PowerPath software provided load balancing and failover for the dual HBAs that were present in each system. All systems used raw devices for the database data files. The use of raw devices removed the overhead of a file system and allowed the database to access and manage the storage directly.

The test team set up the database tablespaces exactly the same on each system (see Figure 3). The two 10-drive RAID-10 LUNs were used for the data, index, undo, and temporary tablespaces. Each tablespace consisted of two data files, with one data file on each of the RAID-10 LUNs. A RAID-1 LUN was used for logs.

Running the application

To study the performance of an online transaction processing (OLTP) database in both scaled-out and scaled-up servers, the test

¹ This term does not connote an actual operating speed of 1 Gbps. For high-speed transmission, connection to a Gigabit Ethernet server and network infrastructure is required.

team simulated the back end of a large online DVD store in a database. The workload exercised the database servers in a common scenario, running end-of-quarter financial rollup reports while processing new orders.

The database comprised a set of data tables organized according to a certain schema, as well as a set of stored procedures that did the actual work of managing the data in the database while orders were entered and reports were requested. The database back end was designed to be driven from a Web-based middle tier. However, because this test scenario focused on the database servers, the back-end stored procedures were driven directly by custom C programs that simulated a Web-based middle tier.

Database schema

The DVD store comprised four main tables and one minor table (see Figure 4). The Customers table was prepopulated with 200 million customers, including one logical partition containing 100 million U.S. customers and a second partition containing 100 million customers from the rest of the world. The Orders table was prepopulated with 10 million orders per month, starting in January 2003 and ending in September 2003, with each month in a separate logical partition. The Orderlines table was prepopulated with an average of five items per order, also partitioned by month. The Products table contained 1 million DVD titles. In addition, the Categories table listed the 16 DVD categories.

Stored procedures

The DVD store database was managed using seven stored procedures. The first two procedures were used during the login phase. If the

A cluster of redundant servers can help administrators improve the availability of business-critical applications over a single large server by eliminating the server as a single point of failure.

customer was a returning customer, Login was used to retrieve the customer's information, in particular the CUSTOMERID. If the customer was a new customer, New_customer was used to create a new row in the Customers table with the customer's data. Following the login phase, the customer might search for a DVD by category, actor, or title using Browse_by_category, Browse_by_actor, or Browse_by_title, respectively. Finally, after the customer had made a selection, Purchase was called to complete the transaction (for this stored procedure, visit *Dell Power Solutions* online at http://www.dell.com/magazines_extras). Additionally, Rollup_by_category was used to total the sales by DVD category for the previous month, quarter, and half-year periods (for this stored procedure, visit http://www.dell.com/magazines_extras).

Driver applications

The test team wrote separate multithreaded driver programs to model the OLTP order-entry workload and the report request workload.

Online transaction processing. Each thread of the OLTP driver application connected to the database and made a series of stored procedure calls simulating users logging in, browsing, and purchasing. Database connections remained full because the OLTP driver simulated no user think times or key times—similar to the behavior of a real multitiered application in which a small number of connections are pooled and shared among Web servers that may be handling thousands of simultaneous customers. In this way, the test simulated database activity realistically without modeling thousands of users.

Each thread of the OLTP driver modeled a series of customers going through the entire sequence of logging in, browsing the catalog using several methods, and finally purchasing selected items. Each completed customer sequence counted as a single order. The driver measured order rates and the average response time to complete each order. Several tunable parameters controlled the application, as described in Figure 5.

Tablespace	Contains	Space used/available
CUSTTBS	Customers table	34 GB/38 GB
INDXTBS	Indexes	30 GB/32 GB
ORDERTBS	Orders and orderlines tables	22 GB/24 GB
DS_MISC	Products and categories tables	0.5 GB/1 GB
UNDOTBS	Undo tablespace	1 GB/3 GB
TEMP	Temporary table	18 GB/20 GB

Figure 3. Database tablespaces used in the test environment

Table	Columns	Number of rows
Customers	CUSTOMERID, FIRSTNAME, LASTNAME, ADDRESS1, ADDRESS2, CITY, STATE, ZIP, COUNTRY, REGION, EMAIL, PHONE, CREDITCARD, CREDITCARDEXPIRATION, USERNAME, PASSWORD, AGE, INCOME, GENDER	200 million
Orders	ORDERID, ORDERDATE, CUSTOMERID, NETAMOUNT, TAX, TOTALAMOUNT	90 million
Orderlines	ORDERLINEID, ORDERID, PROD_ID, QUANTITY, ORDERDATE	450 million
Products	PROD_ID, CATEGORY, TITLE, ACTOR, PRICE, QUAN_IN_STOCK, SPECIAL	1 million
Categories	CATEGORY, CATEGORYNAME	16

Figure 4. Database schema for DVD store scenario used in test environment

Parameter	Description	Value(s) used in test
n_threads	Number of simultaneous connections to the database	See Figure 6
warmup_time	Warm-up time before statistics are kept	1 minute
run_time	Runtime during which statistics are kept	Varied
pct_returning	Percent of customers who are returning	95 percent
pct_new	Percent of customers who are new	5 percent
n_browse_category	Number of searches based on category	Range: 1–3 Average: 2
n_browse_actor	Number of searches based on actor	Range: 1–3 Average: 2
n_browse_title	Number of searches based on title	Range: 1–3 Average: 2
n_line_items	Number of items purchased	Range: 1–9 Average: 5
net_amount	Total amount of purchase	Range: \$0.01–\$400.00 Average: \$200.00

Figure 5. OLTP driver parameters

Reports. The report request driver program was similar to the OLTP driver application in that each thread connected to the database and started making stored procedure calls. Each thread made repeated calls to the Rollup_by_category stored procedure, which calculates total sales by DVD category for the previous month, quarter, and half year until reports for all 16 categories are completed. In each test, eight simultaneous reports were run.

Observing the results

To compare the performance of the two-node Dell cluster against the single IBM eServer xSeries 445 server, the test team ran both OLTP and report request driver programs simultaneously—first against the Dell cluster (with one node handling the OLTP requests and the other handling the reports), then against the IBM server. OLTP orders per minute (where *order* is defined as one complete login, browse, and purchase sequence) and response times, as well as the average completion time of the financial rollup reports, were measured by the two driver programs.

CPU utilization was measured using the Windows Performance Monitor. The test determined how many orders per minute each database server could handle while simultaneously running eight reports and keeping server CPU


System administrators
can cost-effectively
increase the processing
power available to
enterprise applications
by adding more
servers, or nodes, to
scale out a cluster.

	Dell PowerEdge 6650 cluster (two servers, each with four processors)	IBM eServer xSeries 445 (eight processors)
Simultaneous OLTP database connections	18	13
Simultaneous report requests	8	8
OLTP orders per minute	28,790	20,338
OLTP average response time (seconds)	0.033	0.035
Report average completion time (minutes)	17.6	16.0
CPU utilization (minutes)	Node 1: 89.1 Node 2: 89.2	89.9

Figure 6. Test results for Dell PowerEdge 6650 cluster and IBM eServer xSeries 445 server

utilization under 90 percent—a typical system target that allows for order spikes. As shown in Figure 6, the Dell cluster handled 42 percent more orders per minute than the eight-processor IBM server, while taking only 10 percent more time to process the reports.

Gaining expandability and redundancy with clustered servers

Simulating a typical mixed workload for an online store running on a leading enterprise database server, a two-node Dell PowerEdge 6650 cluster with four Intel Xeon processors MP at 2.8 GHz (totaling eight processors) outperformed by 42 percent an IBM eServer xSeries 445 equipped with eight Intel Xeon processors MP at 2.8 GHz. Such cluster performance enables system administrators to lower the cost of computing by scaling out industry-standard servers flexibly and cost-effectively to meet the demands of large-scale enterprise applications. In addition, a cluster of redundant servers can help administrators improve the availability of business-critical applications over a single large server by eliminating the server as a single point of failure. 

Dave Jaffe, Ph.D. (dave_jaffe@dell.com) is a senior consultant on the Dell Technology Showcase team who specializes in cross-platform solutions. Previously, he worked in the Dell Server Performance Lab, where he led the team responsible for Transaction Processing Council (TPC) benchmarks. Before working at Dell, Dave spent 14 years at IBM in semiconductor processing, modeling, and testing, and in server and workstation performance. He has a Ph.D. in Chemistry from the University of California, San Diego, and a B.S. in Chemistry from Yale University.

Todd Muirhead (todd_muirhead@dell.com) is an engineering consultant on the Dell Technology Showcase team. He specializes in SANs and database systems. Todd has a B.A. in Computer Science from the University of North Texas and is Microsoft Certified Systems Engineer + Internet (MCSE +I) certified.

FOR MORE INFORMATION

Dell PowerEdge 6650: http://www1.us.dell.com/content/products/productdetails.aspx/pedge_6650?c=us&cs=555&l=en&s=biz

Optimizing Disaster Recovery

Using Oracle Data Guard on Dell PowerEdge Servers

The high cost of computing system downtime has prompted organizations to view business continuity and high availability as two critical IT concerns. Using Oracle9i™ Real Application Clusters and Oracle® Data Guard on Dell™ PowerEdge™ servers and Dell storage can help administrators cost-effectively achieve data protection, high availability, and resilience for IT infrastructures.

BY PAUL RAD, ZAFAR MAHMOOD, IBRAHIM FASHHO, RAYMOND DUTCHER, LAWRENCE TO, AND ASHISH RAY

Although many formulas exist for calculating the cost of downtime, most IT managers would agree that downtime is simply not acceptable for business today. Enterprises expect their systems to be up and running without interruption, in many cases 24/7. Downtime, whether planned or unplanned, translates into lost opportunities and increased costs.

Dell and Oracle have partnered to offer a high-availability architecture that helps minimize scheduled and unscheduled downtime caused by numerous events, including system failures, site disasters, user errors, data corruption, and maintenance activities.

Dell produces entry-level, midrange, and high-end server and storage clusters built from standards-based components. These clusters are designed to help improve availability by removing all single points of failure within the cluster. At each cluster level, Dell also provides the capability to recover from additional failures, thereby protecting against multiple component failures. Low-cost Intel® processor-based Dell™ PowerEdge™ servers can help businesses implement the degree of availability that best meets their service level objectives.

Oracle® Maximum Availability Architecture (MAA), a High Availability (HA) best-practices blueprint from Oracle, aims to maximize system availability while reducing the

design complexity of an optimal HA architecture. MAA, which uses Oracle HA technologies such as Oracle9i™ Real Application Clusters (RAC) and Oracle Data Guard, provides recommendations that encompass the database tier, the application server tier, network and storage infrastructures, and operational principles. By adopting the MAA methodology, IT organizations can build a simple, robust architecture that helps prevent, detect, and recover from outages with a fast mean time to recovery (MTTR).

Dell and Oracle recommend adopting the MAA methodology on Dell platforms to address requirements for high availability, data protection, and disaster recovery. This article—the result of a joint project between Dell and Oracle engineering teams—explains how the RAC and Data Guard components of the Oracle MAA can be used on Dell PowerEdge servers and storage to build the foundation of an end-to-end, high-availability architecture.

Creating a highly available infrastructure using Oracle9i RAC

In an Oracle9i RAC environment, each node in a cluster runs a separate Oracle instance, and these instances can concurrently access a single, shared database. Although it spans multiple hardware systems, the database appears to applications as a single, unified database system. This

configuration helps provide a very high degree of scalability and availability to enterprise applications:

- The capability to flexibly, transparently, and cost-effectively scale capacity as business needs change
- Fault tolerance to failures within the cluster—particularly node failures

A typical Dell and Oracle9i configuration includes a storage area network (SAN). A Dell/EMC Fibre Channel-based SAN fabric supports multipath routing between SAN switches, helping ensure that no single points of failure exist in the configuration. In a typical topology, a node has multiple Fibre Channel host bus adapters (HBAs), each connected to the same SAN, resulting in multiple paths to the same device. SAN storage devices also can accept multiple Fibre Channel connections.

Although Oracle9i RAC addresses local system failures and provides rapid recovery from node failures or instance crashes, it does not offer protection from site disasters or user errors such as an accidental drop of critical user tables in the database. Such protection is provided by Oracle Data Guard, an integrated feature of the Oracle9i Database Enterprise Edition.

Enabling disaster recovery using Oracle Data Guard

Oracle Data Guard is software that creates, maintains, and monitors one or more standby databases to help protect enterprise data from failures, disasters, user errors, and data corruption. Data Guard maintains standby databases as transactionally consistent copies of the primary database. These standby databases can be located at remote disaster recovery sites thousands of miles from the production data center or they may be located in the same building. If the primary database becomes unavailable because of a planned or unplanned outage, Data Guard can be used to switch any standby database to the primary role, thus minimizing the downtime associated with the outage and helping to prevent data loss.

A standby database is initially created from a backup copy of the primary database. Once the standby database is created, Data Guard automatically maintains it by transmitting primary database redo data to the standby system over standard TCP/IP networks and then applying the redo data to the standby database.

Data Guard supports two types of standby database, each of which uses a different method to apply redo data to the standby

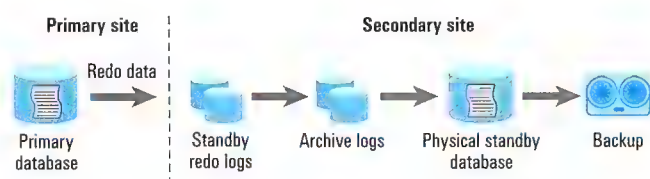


Figure 1. Data Guard Redo Apply (physical standby database)

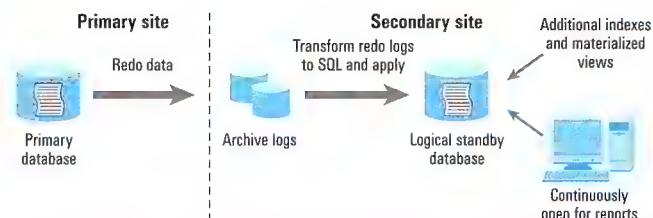


Figure 2. Data Guard SQL Apply (logical standby database)

database and keep it transactionally consistent with the primary database: Redo Apply, used for physical standby databases, and SQL Apply, used for logical standby databases.

- **Redo Apply process on a physical standby database:** A physical standby database is kept synchronized with the primary database by applying the redo data received from the primary database using Oracle media recovery (see Figure 1). The standby database is physically identical to the primary database on a block-for-block basis, and thus the database schemas, including indexes, are the same. The physical standby database can be opened in read-only mode, and queries can be run on it at that time; however, it cannot run recovery at the same time it is opened as read-only.
- **SQL Apply process on a logical standby database:** A logical standby database contains the same logical information as the primary database, but the physical organization and structure of the data may be different. The SQL Apply process keeps the logical standby database synchronized with the primary database by transforming the redo data received from the primary database into SQL statements and then executing the SQL statements on the standby database (see Figure 2). This enables the logical standby database to be accessed for queries and reporting purposes at the same time the SQL statements are being applied to it.

Figure 3 shows two sites with identical configurations. Each site consists of redundant components so that requests can always be serviced, even if a failure occurs. Each site also contains a set of application servers or mid-tier servers.

The primary site with the primary database uses Oracle9i RAC to protect the database from host and instance outages. The secondary site contains a physical standby database that is maintained by the Data Guard Redo Apply process. The secondary site uses Oracle9i RAC to protect it from local host and instance outages.

Oracle Data Guard offers two simple methods—switchover and failover—to handle both planned and unplanned outages of the primary database. Administrators can initiate both methods directly through simple SQL statements or the Data Guard Manager graphical user interface (GUI), which is the Data Guard administrative interface integrated with Oracle Enterprise Manager.

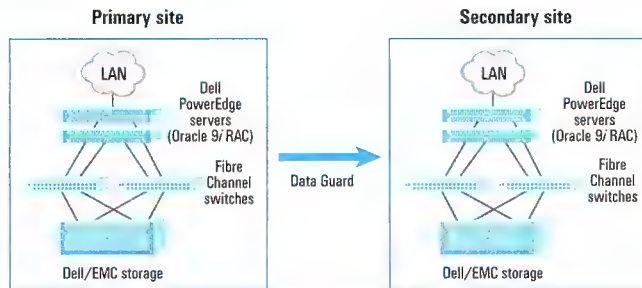


Figure 3. Maximum Availability Architecture using Oracle9i RAC and Oracle Data Guard

Data Guard switchover

Switchover is a planned role reversal of the primary and standby databases to manage scheduled maintenance on the primary database. A switchover operation does not require re-instantiation of the Oracle database, so the primary database can assume the role of a standby database almost immediately. As a result, administrators can perform scheduled maintenance more easily and frequently. For example, administrators can use switchover to perform a system upgrade on the primary site by switching over all the database clients to the standby site as they upgrade hardware on the primary site.

For steps on how to invoke a switchover in a Data Guard configuration, refer to “Physical Standby Database Switchover” in the Oracle Maximum Availability Architecture white paper at http://otn.oracle.com/deploy/availability/pdf/MAA_WP.pdf. For detailed information on how to configure an Oracle RAC Data Guard physical standby database, visit *Dell Power Solutions* online at http://www.dell.com/magazines_extras.

Following a successful switchover, the standby database assumes the primary role and the former primary database becomes a new standby database. In a RAC environment, a switchover requires only one active Oracle instance for each database. If required, administrators may perform a *switchback* operation by doing a subsequent switchover to return the databases to their original roles.


Data Guard failover

Administrators may invoke a *failover* operation when an unplanned catastrophic failure occurs on the primary database or when the primary database cannot be recovered in a timely manner. A Data Guard failover may be accompanied by a site failover to move end users to the new site and database. Once the failover is completed, the primary database can be accessed from the secondary site. Following MAA guidelines, the former primary database must be re-created as a new standby database to restore resiliency.

Typically, little or no data loss is experienced during a failover operation. For detailed information about Data Guard failover, refer to “Physical Standby Database Failover” in the Oracle Maximum Availability Architecture white paper.

Enabling business continuity with Oracle9i RAC and Data Guard

The combination of Oracle MAA, Dell PowerEdge servers, and Dell storage can offer enterprises an easy, low-cost means to implement business continuity and disaster recovery for IT infrastructures. Oracle9i RAC running on Dell PowerEdge servers helps provide the reliability and scalability of a redundant cluster environment. RAC enables high availability by helping provide continuous data access when a node or instance fails, or when performing scheduled system maintenance on a subset of nodes in the cluster.

Oracle Data Guard facilitates data protection and disaster recovery by automating the maintenance of geographically distant standby databases as transactionally consistent copies of the primary database. Data Guard enables easy switchover or failover of a primary database to a standby database if planned or unplanned outages occur at the primary site. Because it is an integrated feature of the Oracle database, Data Guard can be more cost-effective and better optimized to protect Oracle data than host-based or storage-based remote mirroring.¹ For continuous data availability and a resilient, high-availability system architecture, organizations may consider implementing Oracle MAA best practices in combination with Oracle Data Guard and Oracle9i RAC on Dell servers and storage. 

Paul Rad (paul_rad@dell.com) is a senior software engineer in the Dell Database and Application Engineering Department of the Dell Product Group.

Zafar Mahmood (zafar_mahmood@dell.com) is a software engineer in the Dell Database and Application Engineering Department of the Dell Product Group.

Ibrahim Fashho (ibrahim_fashho@dell.com) is the development manager for the Database and Application Engineering Department of the Dell Product Group.

Raymond Dutcher (raymond.dutcher@oracle.com) is a principal member of the technical staff in the Oracle High Availability Systems Group.

Lawrence To (lawrence.to@oracle.com) is a principal member of the technical staff in the Oracle High Availability Systems Group.

Ashish Ray (ashish.ray@oracle.com) is a senior product manager in the Oracle Database High Availability Group.

FOR MORE INFORMATION

Dell and Oracle partnership:

<http://www.dell.com/oracle>

Oracle9i RAC:

<http://otn.oracle.com/products/database/clustering>

Oracle Data Guard:

<http://otn.oracle.com/deploy/availability/htdocs/DataGuardOverview.html>

Oracle Maximum Availability Architecture:

<http://otn.oracle.com/deploy/availability/htdocs/maa.htm>

Oracle Database High Availability:

<http://otn.oracle.com/deploy/availability>

¹For a comparison of Oracle Data Guard and third-party remote mirroring options, visit <http://otn.oracle.com/deploy/availability/htdocs/DGCompTech.html#RemoteMirror>.

Introducing VMware ESX Server, VirtualCenter, and VMotion on Dell PowerEdge Servers

VMware® ESX Server™ software enables administrators to provision multiple independent virtual machines on the same physical server. Dell engineers tested VMware ESX Server, VirtualCenter virtual machine management software, and VMotion™ virtual machine migration technology on Dell™ PowerEdge™ servers to illustrate how virtual machines can be moved from one physical server to another while processing heavy production loads.

BY DAVE JAFFE, PH.D.; TODD MUIRHEAD; AND FELIPE PAYET

IT managers today face a number of challenges as they are pushed to do more with less: improving service delivery, decreasing server sprawl, increasing system utilization, and making IT resources more flexible. VMware® server virtualization software running on Dell™ PowerEdge™ servers can help address these challenges.

VMware ESX Server™ software enables administrators to create multiple *virtual machines* on a single Intel® processor-based server, where each virtual machine can run a separate operating system (OS) and applications. VMware VirtualCenter provides centralized virtual machine monitoring and management from an easy-to-use graphical user interface (GUI). VMotion™ virtual machine migration technology enables administrators to move a running virtual machine from one physical server to another.

The Dell and VMware approach targets specific workload deployments in which server virtualization can offer the most value:

- **Test and development environments:** Virtualization can help administrators consolidate multiple

test and development servers onto fewer physical servers without sacrificing flexibility or functionality.

- **Application consolidation:** Virtualization enables administrators to consolidate applications from underutilized systems onto fewer physical servers, helping to simplify systems management and lower total cost of ownership (TCO) without compromising stability or security.

By deploying VMware ESX Server on multiple two- or four-processor servers and leveraging VMware VMotion, administrators may achieve several benefits not available on deployments that comprise a single server using eight or more processors:

- **Risk mitigation:** Virtual machines distributed among smaller servers can mitigate the impact of a hardware failure. In comparison, the failure of a single larger system would affect all virtual machines hosted by that one server.

Disk enclosures	Disks														
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
First DAE2											RAID-5: data staging				
Second DAE2	RAID-5: virtual machine (VM) boot drives														
Third DAE2	RAID-5: SQL data 1 (VM 1)					RAID-5: SQL data 2 (VM 1)					RAID-1: Logs (VM 1)		RAID-1: SnapView cache		Hot spare
Fourth DAE2	RAID-5: SQL data 1 (VM 2)					RAID-5: SQL data 2 (VM 2)					RAID-1: Logs (VM 2)				

Figure 1. Organization of LUNs on the Dell/EMC CX600 storage array

- **Expansion flexibility:** Deployments based on smaller, industry-standard building blocks permit a modular approach to expandability, whereby organizations can add incremental capacity using two- and four-processor servers instead of eight-processor and larger systems.
- **Operational flexibility:** A VMware deployment based on multiple Dell servers allows the live migration of virtual machines from one physical server to another using VMotion technology. This approach enables administrators to respond quickly to changes in workload demand and perform hardware upgrades or maintenance, all with minimal impact to workload delivery.

To evaluate Dell servers as a platform for server virtualization, in December 2003 a team of Dell engineers tested the performance of ESX Server software on the four-processor PowerEdge 6650 server using a Dell/EMC® storage area network (SAN). The Dell test team built an application that models an online DVD store on two instances of Microsoft® SQL Server™ 2000 Enterprise Edition. These database instances were deployed as virtual machines on two separate PowerEdge 6650 servers. One database instance received orders, and the other generated financial reports based on the order data. To determine whether virtual machines running heavy loads in a production environment could be moved without service interruption, the Dell team moved the virtual machine hosting the order entry database from one physical server to the other while the database was processing 100 orders per second—with no loss of transactions and only a slight rise in response time.

Setting up the hardware for the test environment

The two 4U Dell PowerEdge 6650 servers were each configured with VMware ESX Server 2.0.1 and four Intel® Xeon™ processors MP at 2.8 GHz with 2 MB of level 3 (L3) cache and 4 GB of RAM. Each PowerEdge 6650 server used a PowerEdge Expandable RAID Controller 3, Dual Channel (PERC 3/DC) and an Intel PRO/1000XT Gigabit Ethernet¹ network interface card (NIC) in addition to two on-board Gigabit Ethernet NICs. The

three NICs allowed dedicated bandwidth for the ESX Server service console, the virtual machines, and the VMotion workload management software.

The PowerEdge 6650 servers were attached to the Dell/EMC SAN by a QLogic® 2340 Fibre Channel host bus adapter (HBA). A Dell/EMC CX600 storage array was attached to the SAN to provide shared storage. The test team assigned 38 of the 150 drives attached to the CX600 for the VMware environment. The basic configuration of the CX600 storage array was as follows:

- **Disk enclosures:** Four Dell/EMC DAE2 disk array enclosures
- **Disks:** Thirty-eight 73 GB disks at 10,000 rpm
- **Logical storage units (LUNs):** One 6-disk RAID-5 LUN for the virtual machine boot drive, four 5-disk RAID-5 LUNs for database data storage, two 2-disk RAID-1 LUNs for database logs, one 5-disk RAID-5 LUN for temporary data staging before loading, one 2-disk RAID-1 split into two LUNs for an EMC SnapView™ storage management software cache, and one hot-spare disk
- **Software:** EMC Navisphere® Manager, EMC Access Logix™, and SnapView

One LUN on the Dell/EMC CX600 array was used to stage the data that was loaded into the database (see Figure 1). Using the snapshot capability of the SnapView software, the test team created a second copy of this data so that both virtual machines could load the data simultaneously.

Dell engineers used a PowerEdge 2650 server to produce a transaction load to run against the databases that were installed in the virtual machines on the two PowerEdge 6650 servers (see Figure 2). All servers, including the PowerEdge 6650 servers, were connected to a Dell PowerConnect™ 5224 Gigabit Ethernet switch for network connectivity. Using a Brocade® Fibre Channel switch, the test team also attached the PowerEdge 6650 servers to the Dell/EMC CX600 storage array.

All storage for the ESX Server-based virtual machines resided on the SAN, and each virtual machine was configured with its own

¹ This term does not connote an actual operating speed of 1 Gbps. For high-speed transmission, connection to a Gigabit Ethernet server and network infrastructure is required.

boot drive as well as two data drives and one log drive. When testers moved a virtual machine from one physical server to the other, only the RAM contents of the migrating virtual machine moved with it to the new physical hardware. Both servers already had access to storage, which was shared on the SAN.

Setting up the software for the test environment

The two VMware software products used for the Dell test were ESX Server and VirtualCenter. ESX Server has its own kernel that runs directly on the hardware and hosts virtual machines, enabling multiple virtual machines to run at the same time on the same hardware. VirtualCenter is a console application through which administrators can monitor and control ESX Server installations—and the virtual machines running on them—from a central location across multiple Dell servers.

Installing and configuring ESX Server

The test team configured the internal drives on the PowerEdge 6650 servers as RAID-1. The QLogic HBA was disconnected from the SAN during the initial stage of the ESX Server installation. To install ESX Server, Dell engineers booted from the ESX Server CD and answered the installation questions concerning partitioning of the local drives, the ESX Server host name, IP address, Domain Name System (DNS) server, gateway address, and initial root password. The team copied all necessary files from the installation CD and then rebooted the system.

To complete the installation of the ESX Server software—and for most administration and configuration tasks—the team accessed the ESX Server service console remotely through a Web browser. Following the initial installation stage, when administrators access the ESX Server for the first time through a Web browser, the software presents a series of configuration steps. These steps include installing the ESX Server license and configuring all hardware on the server that would be used by either the service console or virtual machines. The service console portion of each ESX Server installation requires a dedicated NIC. Dell recommends that the virtual machines also use one or more dedicated NICs per physical server. In this test, each virtual machine controlled its own HBA and all the SAN storage allocated to it.

After configuring the ESX Server hardware options, administrators must reboot the server. Just prior to rebooting, the Dell team connected the QLogic HBA into the SAN fabric and created a new zone on the switch for the newly connected server. Once the switch was correctly zoned, Dell engineers used Navisphere Manager—the management tool for the Dell/EMC CX600 storage array—to manually register the new host in the Connectivity Status screen. (Currently, no version of Navisphere Agent is available to register ESX Server automatically.) Once the registration was complete, the team used Navisphere Manager to create the necessary RAID groups, LUNs,

and storage groups. Administrators must assign all ESX Server systems expected to participate in VMotion virtual machine migrations to the same storage group.

Adding ESX Server servers to the VirtualCenter service console

VirtualCenter, a Microsoft Windows®-based program, was installed on a separate PowerEdge 1750 system that served as the management node for the test configuration. Dell engineers added all ESX Server virtual machines to be managed by VirtualCenter to the VirtualCenter console using a simple Connect Host wizard, which prompted for the host name, user ID, and password of each system running ESX Server. Adding all the ESX Server systems to VirtualCenter enables administrators to perform management tasks—including cloning, template production, and VMotion virtual machine migration—from the VirtualCenter console for any virtual machines that reside on those ESX Server systems.

Creating the virtual machines

The Dell team used VirtualCenter to create a new virtual machine on the SAN, specifying the Microsoft Windows Server™ 2003 Enterprise Edition OS, a 10 GB hard disk, 1 GB of RAM, and two CPUs. (The symmetric multiprocessing, or SMP, feature of ESX Server allowed the virtual machine to use two physical CPUs.) VirtualCenter created a virtual machine ready for installation of the OS. The Dell team then booted the virtual machine from the ISO image of the Windows Server 2003 Enterprise Edition installation CD and installed the OS on the virtual machine. The database application was installed afterward. Dell engineers then created two clones of this virtual machine

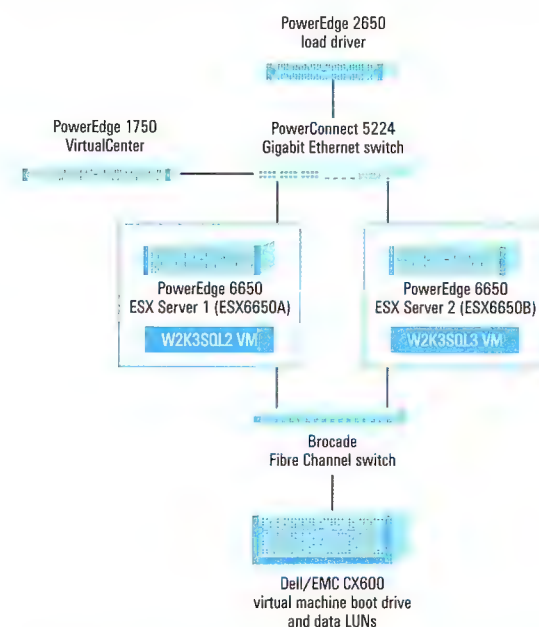


Figure 2. Configuration of servers and storage used for testing

master for use in testing. After the virtual machines were created, each was assigned additional hard disks for the data and logs of the database that resided on the CX600 storage array (see “Setting up the hardware for the test environment”).

Examining the test database application: An online DVD store

To demonstrate the advantages of running a large application on VMware ESX Server, Dell engineers created a 100 GB online DVD store, which they implemented as two replicated database instances, each running on its own virtual machine. One of the database instances handled the entry of new orders and replicated changes on a scheduled basis to the second database instance, which was used for generating financial reports.

The DVD store database consisted of a set of data tables organized according to a certain schema, as well as a set of stored procedures that did the actual work of managing the data in the database as orders were entered and reports requested. The database back end was designed to be driven from a Web-based middle tier, but because the focus of the Dell test was on the database servers, the back-end stored procedures were driven directly by custom programs written in the C programming language to simulate a Web-based middle tier.

Understanding the database schema

The DVD store comprised four main tables and one additional table (see Figure 3). The Customers table was prepopulated with 200 million customers: 100 million U.S.-based customers and 100 million customers from the rest of the world. The Orders table was prepopulated with 10 million orders per month, starting in January 2003 and ending in September 2003. The Orderlines table was prepopulated with an average of five items per order. The Products table contained 1 million DVD titles. An additional Categories table listed the 16 DVD categories. For the full DVD store database build script used in this test, visit *Dell Power Solutions* online at http://www.dell.com/magazines_extras.

Managing the database using stored procedures

The Dell team managed the DVD store database using seven stored procedures. The first two procedures were used during the login phase. For returning customers, the Login procedure retrieved the customer's information—in particular, the CUSTOMERID. For new customers, the New_customer procedure created a new row in the Customers table containing the customer's data.

Following the login phase, the customer might search for a DVD by category, actor, or title. These database functions were implemented by the Browse_by_category, Browse_by_actor, and Browse_by_title procedures, respectively. Finally, after the customer completed the selections, the Purchase procedure was called to complete the transaction. Additionally, the Rollup_by_category procedure calculated total sales by DVD category for the previous month,

Table	Columns	Number of rows
Customers	CUSTOMERID, FIRSTNAME, LASTNAME, ADDRESS1, ADDRESS2, CITY, STATE, ZIP, COUNTRY, REGION, EMAIL, PHONE, CREDITCARD, CREDITCARDEXPIRATION, USERNAME, PASSWORD, AGE, INCOME, GENDER	200 million
Orders	ORDERID, ORDERDATE, CUSTOMERID, NETAMOUNT, TAX, TOTALAMOUNT	90 million
Orderlines	ORDERLINEID, ORDERID, PROD_ID, QUANTITY, ORDERDATE	450 million
Products	PROD_ID, CATEGORY, TITLE, ACTOR, PRICE, QUAN_IN_STOCK, SPECIAL	1 million
Categories	CATEGORY, CATEGORYNAME	16

Figure 3. Database schema for online DVD store

quarter, and half-year periods. For the stored procedures, visit *Dell Power Solutions* online at http://www.dell.com/magazines_extras.

Using driver programs to model workloads

The Dell team wrote separate multithreaded driver programs to model the order entry, or *online transaction processing* (OLTP), workload as well as the report request workload. Each thread of the OLTP driver application connected to the database and made a series of stored procedure calls that simulated customers logging in, browsing, and purchasing. Because Dell engineers did not simulate customer think time or key time, the database connections were kept full—simulating a multitiered application in which a few connections are pooled and shared among Web servers that may be handling thousands of simultaneous customers. In this way, the test team achieved a realistic simulation of database activity without having to model thousands of customers.

Each thread of the OLTP driver program modeled a series of customers going through the entire sequence of logging in, browsing the catalog several ways, and purchasing selected items. Each completed customer sequence counted as a single order. The OLTP driver program measured order rates and the average response time to complete each order. Several tunable parameters were used to control the application (see Figure 4).

The report request driver program was similar to the OLTP driver program in that each thread connected to the database and started making stored procedure calls. Each thread made repeated calls to the Rollup_by_category stored procedure until reports for all 16 DVD categories were completed. In each test, eight simultaneous reports were run.

Moving a virtual machine under heavy load

To demonstrate the capability of VMware software to move virtual servers around a farm of physical servers, the Dell team used the VMware VMotion add-on to VirtualCenter, which enables administrators to move a virtual machine from one physical server running

ESX Server to another. The migration was performed while the virtual machine was running the DVD store database under a heavy stress load of 100 orders per second. In a live production environment, such a move might be required to balance workloads among computing resources, perform routine maintenance on a server, or respond to an alert that a server parameter such as temperature had exceeded a warning threshold. In Figure 5, the VirtualCenter console shows the virtual machines in the test server farm.

At the start of the test, one node of the database replication group, W2K3SQL3, on physical PowerEdge 6650 server ESX6650B was handling approximately 100 orders per second with an average response time of 0.1 second. For the test, response time was defined as the total response time experienced by the simulated customer for the complete order transaction, including login time, browse time, and response time after the customer pressed the Submit button to purchase the order.

Dell engineers then started the second database system, W2K3SQL2, running on physical server ESX6650A, which began calculating sales by DVD category for eight separate categories. In addition, the test team set up the servers to replicate new orders from the W2K3SQL3 node to the W2K3SQL2 node once per day. The two virtual machines running database instances are shown in the ESX Server service console in Figure 6.

The Dell team started the order entry and the report request workloads against the two database instances, each instance running in a virtual machine on its own PowerEdge 6650 server. Each server achieved full speed—100 orders per minute, or eight simultaneous reports—using about 80 percent of the two CPUs dedicated to each virtual machine. The Dell team used the VMotion

Parameter	Description	Value(s) used in test
n_threads	Number of simultaneous connections to the database	10
warmup_time	Warm-up time before statistics are kept	1 minute
run_time	Runtime during which statistics are kept	Varied
pct_returning	Percent of customers who are returning	95 percent
pct_new	Percent of customers who are new	5 percent
n_browse_category	Number of searches based on category	Range: 1–3 Average: 2
n_browse_actor	Number of searches based on actor	Range: 1–3 Average: 2
n_browse_title	Number of searches based on title	Range: 1–3 Average: 2
n_line_items	Number of items purchased	Range: 1–9 Average: 5
net_amount	Total amount of purchase	Range: \$0.01–\$400.00 Average: \$200.00

Figure 4. OLTP driver parameters

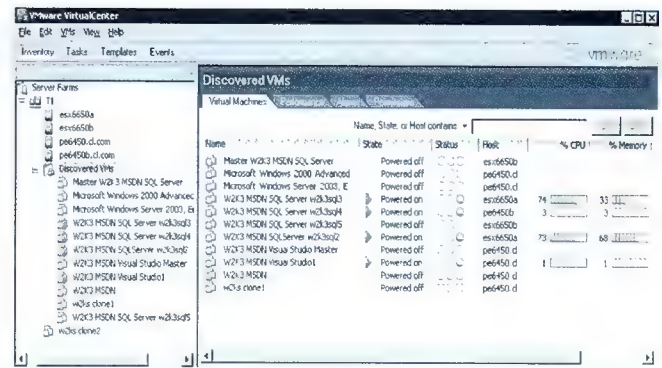


Figure 5. VMware VirtualCenter displaying the virtual machines in the test server farm

feature of VirtualCenter to move the virtual machine performing order entry (W2K3SQL3) from physical server ESX6650B to physical server ESX6650A, without stopping either the incoming orders or the sales calculations on W2K3SQL2. Figures 7 and 8 show the results of this migration.

As shown in Figure 8, for the first 25 seconds after the VMotion migration was initiated at 15:36:20, there was little impact on either throughput (orders per second, indicated in the top half of Figure 8) or response time (indicated in the bottom half of Figure 8) while VirtualCenter prepared for the move by initializing a new virtual machine on the target ESX Server and synchronizing the memory between the two. At about 15:36:45, the effects of the memory synchronization could be seen in the dropping throughput and increasing response time.

The actual move occurred at 15:37:08, and the response time reached a maximum of 2.572 seconds while the order handling paused for approximately two seconds. Immediately after the move, the throughput and response time rapidly returned to close to their previous levels. The target ESX Server CPU

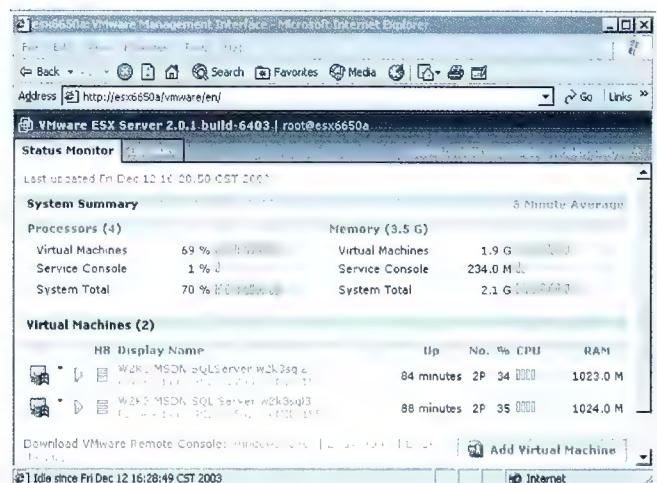


Figure 6. ESX Server service console showing the two virtual machines running database instances

Time	New orders completed per second	Average response time (seconds)	Maximum response time (seconds)
Before VMotion migration	103	0.098	0.201
During VMotion migration	80	0.139	2.572
After VMotion migration	93	0.109	0.492


Figure 7. Performance results before, during, and after VMotion migration of virtual machine running database application under heavy load

utilization rose to about 80 percent as both virtual machines ran on the target server, using two CPUs each. The throughput decreased slightly from the pre-VMotion level but was still high enough to handle 300,000 orders per hour while the first system was being repaired or upgraded.

Using virtual machine migrations to increase operational flexibility

The test findings in this article indicate that ESX Server software running on Dell PowerEdge servers with Dell/EMC SAN storage can provide a robust platform for server virtualization. In the Dell test discussed in this article, two new virtual machines were rapidly cloned from a single master and then used to implement a large online DVD store with one server handling new orders and then replicating the orders to the second server for reporting.

Using VMware VMotion workload management software, an add-on to VMware VirtualCenter, testers demonstrated that a virtual machine handling 100 orders per second could be moved from one physical server to another in less than a minute without stopping the database application and without losing any transactions. Test findings indicate that the slight increase in response time would be nearly imperceptible to the end user. Although the virtual machine migration took 48 seconds, the increased response time of less than three seconds from the end user's perspective—at the point when the virtual machine actually switched from one physical server to the other—was experienced for only a second or two.

Deploying virtual machines on Intel processor-based servers can help IT organizations scale out cost-effectively and respond quickly and flexibly to changes in workload demand. The virtual server approach to IT management also can provide a convenient way to upgrade and maintain production servers in real time, without interrupting service to business-critical applications. In addition, VMware virtual machines running on industry-standard Dell servers can improve system availability and fault tolerance by avoiding a single point of hardware failure, as opposed to a single larger server. 

Acknowledgments

The authors would like to thank Craig Lowery, Tim Abels, and Wenlong Xu of the Scalable Enterprise Computing team at Dell for valuable discussions.

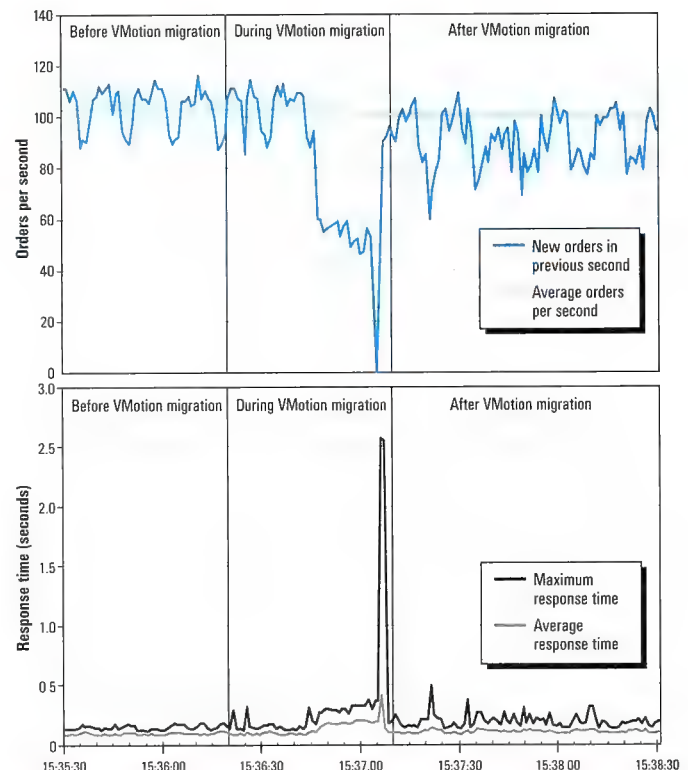


Figure 8. Throughput and response times before, during, and after VMotion migration

Dave Jaffe, Ph.D. (dave_jaffe@dell.com) is a senior consultant on the Dell Technology Showcase team who specializes in cross-platform solutions. Previously, Dave worked in the Dell Server Performance Lab, where he led the team responsible for Transaction Processing Council (TPC) benchmarks. Before working at Dell, Dave spent 14 years at IBM in semiconductor processing, modeling, and testing, and in server and workstation performance. Dave has a Ph.D. in Chemistry from the University of California, San Diego, and a B.S. in Chemistry from Yale University.

Todd Muirhead (todd_muirhead@dell.com) is an engineering consultant on the Dell Technology Showcase team. He specializes in SANs and database systems. Todd has a B.A. in Computer Science from the University of North Texas and is Microsoft Certified Systems Engineer + Internet (MCSE+I) certified.

Felipe Payet (felipe_payet@dell.com) manages the Dell and VMware relationship within the Software Alliance Team of the Dell Enterprise Server Group. Previously, he worked in various product management, business development, and emerging technology marketing roles at Dell, Intel, and several start-ups. Felipe has a B.A. in Economics from Yale University and an M.B.A. from the Sloan School of Management at M.I.T.

FOR MORE INFORMATION

Dell and VMware:
<http://www.dell.com/vmware>

Deploying Dell OpenManage on VMware ESX Server

VMware[™] ESX Server[™] software divides a physical server into a group of logical computing resources by creating multiple, independent partitions—or *virtual machines*—that can run different applications and operating systems on the same hardware platform. Dell[™] OpenManage[™] systems management software works with VMware ESX Server to help maximize uptime of physical servers running virtual machines.

BY BALASUBRAMANIAN CHANDRASEKARAN, TIM ABELS, ROBERT WILSON, AND PAUL RAD

VMware[®] ESX Server[™] software brings virtual computing to Intel[®] processor-based servers. Through logical partitioning, VMware ESX Server creates a group of logical computing resources by dividing one physical server into several independent partitions, or *virtual machines*. These virtual machines reside on the same physical hardware, but operate in isolation from one another. Each virtual machine runs its own separate operating system (OS) instance and one or more applications.

By consolidating applications onto fewer physical servers, administrators can simplify systems management and help lower maintenance and operating costs—increasing the utilization of physical servers to process additional workloads within existing data center facilities. However, the hardware failure of a server that is running multiple applications on multiple virtual machines can be costly and time-consuming. The larger the number of systems affected by a failure, the greater the potential for disruption to business. Proactive monitoring of system health and cyclic system maintenance can be instrumental in preventing unplanned downtime and increasing overall system availability.

The Dell[™] OpenManage[™] product suite offers administrators several systems management capabilities for Dell PowerEdge[™] servers running ESX Server 2.0.1. IT departments can use Dell OpenManage, including Dell

OpenManage Server Administrator, to monitor, manage, and remotely control Dell servers—thereby helping to improve server uptime and lower hardware failure rates. Server Administrator enables IT organizations to track the health of physical servers and detect problematic components before hardware failures occur (see Figure 1).

During development of VMware ESX Server 2.0.1, Dell performed comprehensive testing of its Dell OpenManage products on PowerEdge 6650 servers to help ensure full compatibility of the Dell OpenManage product suite with ESX Server. Dell also provided prerelease versions of Dell OpenManage 3.6 tools and ESX Server 2.0.1 to beta test sites. The beta testing program assisted Dell customers in deploying ESX Server and provided feedback that helped Dell enhance the hardware-virtualization capabilities of its hardware and software products.

Using ESX Server to achieve virtual partitioning

Hardware virtualization through the use of ESX Server software allows multiple OS instances to run simultaneously on the same physical server (see Figure 2). This capability enables administrators to consolidate heterogeneous workloads onto a single physical server by allowing applications that cannot coexist within one OS instance to run on separate OS instances. Note that not all applications are good candidates for virtualization; for example, systems that

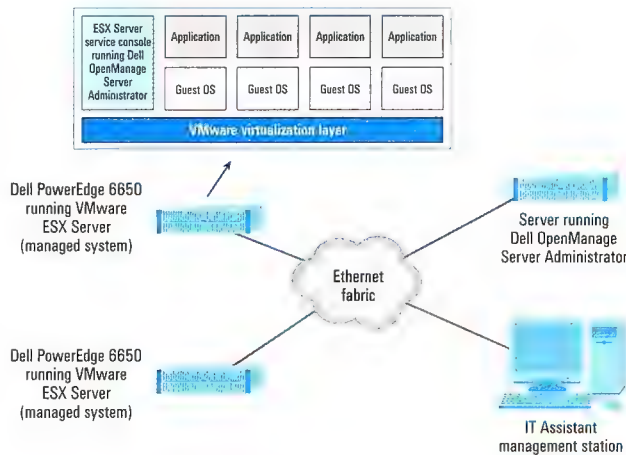


Figure 1. Dell OpenManage deployed on PowerEdge servers running VMware ESX Server

need direct access to physical hardware or those that fully utilize the computing power of the physical hardware should not be implemented on virtual machines. However, hardware virtualization can be a powerful tool for consolidating Web servers and offloading database servers, and for running applications that have moderate demands for memory and CPU resources. By consolidating applications onto fewer physical servers and managing multiple virtual machines through a single ESX Server service console, administrators can streamline operations and increase application availability.

ESX Server encapsulates each virtual machine in a discrete set of files and runs the machine in its own separate environment, thus providing the necessary isolation for running multiple, sometimes incompatible, applications on the same physical hardware. Acting as the *host OS*, ESX Server runs directly on the system hardware, providing powerful resource management features to enable efficient, high-performance server virtualization.

The OS running within a virtual machine is called the *guest OS*, and each virtual machine presents its OS with a consistent set of virtual hardware, regardless of the underlying physical hardware. This hardware independence ensures that only a particular OS instance is affected when an application running within a guest OS becomes unstable and causes that guest OS to crash.

Hardware independence also lets administrators easily relocate virtual machines onto various Intel processor-based servers, even if the physical servers use different underlying hardware. VMware VirtualCenter software can further simplify virtual machine migrations by enabling a *virtual infrastructure* approach to IT management, which allows administrators to move a virtual machine from one physical server to another physical server connected to the same storage area network (SAN) without incurring downtime. (See "Introducing VMware ESX Server, VirtualCenter, and VMotion on Dell PowerEdge Servers" in *Dell Power Solutions*, March 2004.)

Running Dell OpenManage to enhance systems management capabilities

Dell OpenManage provides administrators with comprehensive, one-to-one systems management capabilities within the data center. Dell OpenManage features include proactive monitoring of server health, diagnostics for troubleshooting, alerts and notifications, and remote access. Each server managed by Dell OpenManage software is known as a managed system. Managed-system applications include Dell OpenManage Server Administrator and remote access controller software.

Comprehensive monitoring using Server Administrator

Dell OpenManage Server Administrator provides a browser-based graphical user interface (GUI) that offers a consolidated and consistent way to monitor, configure, update, and manage individual Dell servers. Server Administrator provides the following features:

- Security management
- Command-line interface (CLI)
- Extensive logging
- Diagnostic tools to isolate hardware problems while a server is running
- Remote access to an inoperable server
- Remote administration of a monitored server, including BIOS setup, shutdown, startup, and Dell Remote Access Card III (DRAC III) security

Enhanced availability using DRAC III

DRAC III, a physical card that includes software components, provides alert messages for system problems and also enables remote systems management. Remote management reduces the need for administrators to access servers physically and improves availability by enabling IT departments to manage servers more quickly and address problems proactively before they worsen.

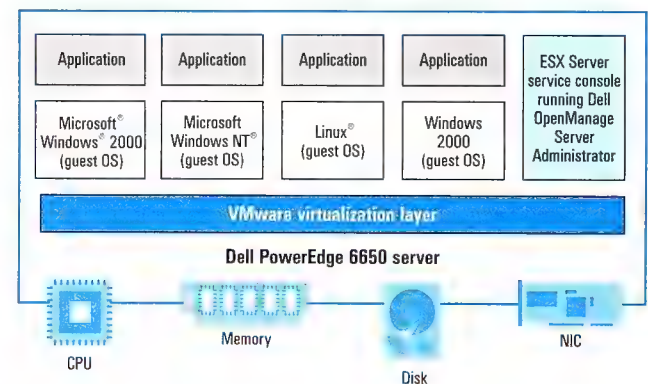


Figure 2. Hardware virtualization using ESX Server

Effective monitoring of managed systems using IT Assistant

A management station can be used to remotely manage one or more servers from a central location. Dell OpenManage IT Assistant is a systems management console program that can be installed on one management station to collect information and provide a view of all managed systems. Server Administrator provides thorough Simple Network Management Protocol (SNMP) integration with IT Assistant as well as third-party management-station programs.

By installing IT Assistant on a management station, organizations can administer thousands of managed systems effectively. IT Assistant provides fault monitoring as well as inventory and asset reporting:

- **Fault monitoring:** Notifications through e-mail, paging, SNMP, or console alerts to keep administrators informed of events concerning disk, memory, voltage, fan, and thermal conditions
- **Inventory and asset reporting:** Service tag number; cost of ownership information; and specifics of the BIOS, micro-processor(s), and memory

Deploying Server Administrator on the ESX Server service console

The ESX Server service console provides a Web-based GUI for managing and configuring virtual machines. Server Administrator, which runs on the service console, can be used to manage the physical server. Note that systems management applications such as Server Administrator and remote access controller software cannot be used within the virtual machines because the software would recognize only the virtualized hardware, not the physical hardware.

Currently, applications running in the service console can monitor only hardware that is either dedicated to the service console or shared with it. For example, Server Administrator can provide information about the physical server as well as subcomponents (such as network components or storage devices) that are assigned to or shared with the service console.

The step-by-step deployment process for Server Administrator on the ESX Server service console is as follows:

1. Install the ESX Server 2.0.1 kernel source from the ESX Server CD and run the Dell OpenManage setup script using the following commands:

```
$ cd /mnt/cdrom/Vmware/RPMS
$ rpm -Uvh kernel-source-2.4.9.-34.i386.rpm
$ /usr/sbin/dellomasetup.pl
```

2. Download ppp-2.4.1-3.i386.rpm from the URL <http://www.vmware.com/download/esx/esx201-openmanage.html>, and install the download using the following command:

```
$ rpm -Uvh ppp 2.4.1-2.i386.rpm
```

3. Configure the SNMP daemon (snmpd) for sending traps to the management station by adding the line:

```
$ trapsink management_station_IP_Address public
```

to the /etc/snmp/snmpd.conf file.

4. Install Dell OpenManage from the Dell Systems Management CD using the following command, and then reboot the system:

```
/mnt/cdrom/start.sh -license
```

Obtaining maximum benefits from server virtualization

Now functioning on Intel processor-based systems, virtual servers can provide significant data center benefits: reduced costs for maintenance, power, and cooling; smaller footprint for server hardware; and simplified systems management processes. As enterprises of all sizes demand higher availability for business-critical applications, the server virtualization approach can help administrators achieve definable efficiencies and cost savings. The combination of industry-standard Dell PowerEdge servers, VMware ESX Server virtualization software, and powerful systems management tools such as the Dell OpenManage suite can enable data center administrators to maximize application availability efficiently and cost-effectively. ➤

Balasubramanian Chandrasekaran (balasubramanian_chan@dell.com) is a systems engineer at the Scalable Enterprise Computing Lab at Dell. His research interests include virtualization of data centers, high-speed interconnects, and high-performance computing. Balasubramanian has an M.S. in Computer Science from The Ohio State University.

Tim Abels (tim_abels@dell.com) is a senior software architect currently developing scalable enterprise computing systems. Tim has an M.S. in Computer Science from Purdue University.

Robert Wilson (rwilson@aig.com) is the manager of the Server Consolidation Group for AIG Technologies, Inc. He has been running production ESX Server systems on Dell PowerEdge servers for more than one year. Robert also has worked with Dell and VMware teams during the Dell/VMware ESX Server 2.0.1 beta program.

Paul Rad (paul_rad@dell.com) is a senior software engineer in the Dell Database and Application Engineering Department of the Dell Product Group. Paul has master's degrees in both Computer Science and Computer Engineering from The University of Texas at San Antonio.

FOR MORE INFORMATION

<http://www.dell.com/vmware>
<http://www.vmware.com>

Implementing Avocent AMX KVM Switches

in the Dell Enterprise Solutions Engineering Group Lab

The Dell™ Enterprise Solutions Engineering Group recently consolidated its labs to enable engineers on different functional teams to share desktops, servers, and storage systems. Dell keyboard, video, mouse (KVM) switches and the Avocent™ AMX™ family of KVM switching products facilitate resource sharing and remote access to the hardware.

BY MIKE KOSACEK AND AVOCENT CORPORATION

The Dell™ Enterprise Solutions Engineering Group recently consolidated several smaller labs into one larger lab facility. The motivation for the move was to allow more efficient sharing of hardware resources across several functional teams—each responsible for the development of certain Dell products including operating systems, database environments, high-availability clusters, and high-performance computing clusters.

To achieve this goal, Dell required flexible keyboard, video, mouse (KVM) switching products that would facilitate local or remote access and allow for future growth. Although each team in the Dell Enterprise Solutions Engineering Group had unique development goals, the teams shared some common lab requirements:

- Many-to-one and any-to-any hardware access
- Remote access to systems on an isolated network
- Sound-isolated work environment and secure data center space for servers and storage racks
- Less-bulky cables
- Direct physical access to systems for multiple engineers (in addition to remote administration over the lab network)

Before consolidating, the Enterprise Solutions Engineering Group labs were using different types of KVM

products from several vendors. All existing KVM switches used analog signaling; most used heavy, bulky cabling that was limited to distances of 20 feet (6 meters) or less from the switch to the server, because analog signals can degrade quickly over long distances. Keyboard, monitor, and mouse connections also were limited in distance because they were required to plug directly into each KVM switch. These factors restricted access and prevented simultaneous systems sharing.

Understanding KVM configuration issues

KVM switches connect multiple systems to a common console that consists of a keyboard, monitor, and mouse. Traditionally, KVM devices have been used for local connectivity in a server rack or within a network operation center or data center. Today, KVM switch capability can be extended across the globe through remote connections using technology such as the Avocent KVM over IP protocol. This protocol enables administrators to accomplish systems management, troubleshooting, and restart tasks even while a system completes its power-on self-test. KVM switches provide an availability advantage over software-based remote administration products, which require the operating system to be running—and which may drop the connection from the client to the server if the client crashes.

Overcoming KVM cabling limitations

Workstations, desktops, and servers use analog signaling for their KVM inputs and outputs. To exceed the 20-foot distance limitation, some KVM switches employ an interface board that amplifies the analog signal so that it can travel greater distances. Some KVM devices also offer analog-to-digital conversion, which enables signals to travel long distances over IP without loss; however, even in these devices, the physical connection from the system to the KVM switch remains analog.

The type of cabling used to connect each system to a KVM switch can vary greatly. Some older switches use heavily-shielded proprietary cabling—either a single large, bulky cable or bundles of cables—to connect keyboard, monitor, and mouse signals. Many KVM devices use industry-standard unshielded twisted-pair (UTP) cabling: the same type of wiring that connects Ethernet network nodes. Modern switches, such as the Avocent® AMX™ series of analog KVM switches, use the standard Category 5 (Cat 5) twisted-pair cabling and RJ-45 connectors. Category 5 Enhanced (Cat 5e) and Category 6 (Cat 6) cabling also can be used. Such cabling enables administrators to reduce the volume of cables significantly—a considerable space savings for racks full of 1U or 2U servers. Standard cables also allow for greater efficiency because most data centers are already wired for Ethernet.

Supporting multiple console connections to KVM switches

Although many KVM switches are limited to one console, some can access multiple consoles. Such switches are often used in network operation centers, where multiple administrators need to work on different systems that reside in the same data center or even in the same rack. If two administrators attempt to simultaneously access the same system—or to access two different systems on the switch—one administrator may be blocked. Depending on the product, a switch may either display a “System is busy” message or allow the blocked administrator view-only access to the system.

Some KVM products allow administrators to connect expansion modules or existing KVM switches to each server port, which increases the number of systems that can be controlled through a single KVM switch; some KVM products also allow administrators to build larger KVM matrices by cascading several switches. When expanding a KVM product or building KVM matrices, administrators should consider how many consoles need simultaneous access to all the systems that are connected to the KVM matrix.

For example, a blocking scenario can occur when two users attempt to access the same system or two different systems in a cascaded switch without enough user connections. Figure 1 shows a KVM switch that provides two console connections and support for 16 systems (a 2 × 16 switch). This switch is cascaded down to a KVM switch with one console connection and support for eight

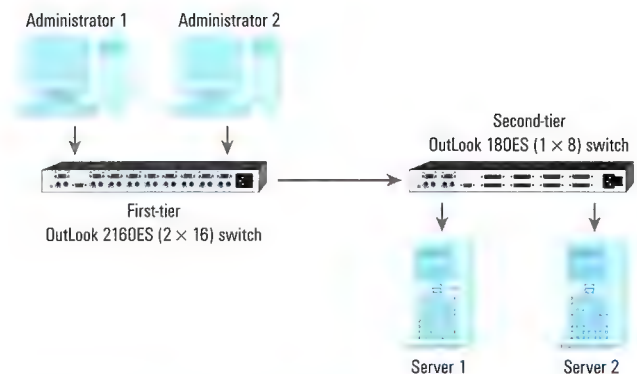


Figure 1. Blocking configuration for KVM switches

systems (a 1 × 8 switch). If administrator 1 is connected to a server in the second tier (server 1), then administrator 2 will be blocked—unable to access any systems in the second tier. This scenario is known as a blocking configuration. If administrator 2 needs to access server 2 in the second tier while administrator 1 is accessing server 1, then the second-tier KVM switch requires at least two console connections to the first-tier switch to ensure a non-blocking configuration.

Implementing KVM switching at Dell

The Dell Enterprise Solutions Engineering Group deployed a combination of Avocent AMX5010 and AMX5000 switches to create the first and second tier of system access, respectively (see Figure 2). Most departments also cascaded the AMX switches to Avocent OutLook® 2160ES switches, allowing the use of a console attached directly to the 2160ES switch as well as a console connection through the AMX switch matrix (by connecting an Avocent AMIQ module to the second output of the 2160ES). To provide an interface from the AMX switches to the Dell 2161DS Remote Console Switch, up to two Avocent AMX5100 user stations were required between the 2161DS and the AMX5010 switch.

The implementation of a 2161DS switch allowed Dell to enable remote KVM sessions. Because the 2161DS switch has both local analog connections and two simultaneous KVM over IP digital connections, it offers the advantage of both local console connections and remote digital KVM sessions over IP. KVM over IP connections enable engineers in the Dell Enterprise Solutions Engineering Group to run tests or check server status from any location from which they can establish a virtual private network (VPN) connection—including over wireless connections established at Dell, from home, or on the road. Remote digital connections also enable remote training or demo sessions to be conducted on servers that reside in the Dell lab.

Administrators implement KVM switches at Dell to supplement several other methods of systems management, including protocols or applications such as Telnet, Microsoft® Terminal Services (in Remote

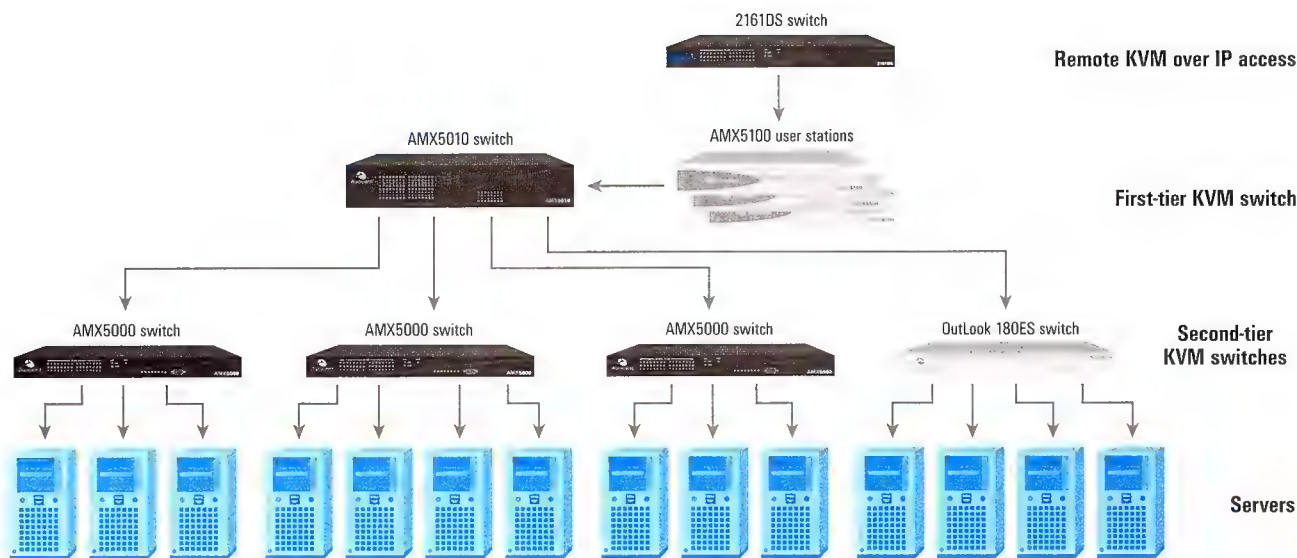


Figure 2. KVM switch deployment at Dell lab

Desktop for Administration mode), and Virtual Network Computing (VNC)—as well as products such as Dell Embedded Remote Access (ERA) and Dell remote access controllers (RACs). KVM switches do not replace systems management tools such as Dell OpenManage™ IT Assistant. Because the development labs at Dell are isolated from the Dell corporate network, the use of these and other administration tools that require network access are limited to use inside the lab.

Setting up cabling and attaching AMIQ modules to AMX switches

Deployment of the Avocent KVM switches at Dell was very straightforward. When planning the new Dell Enterprise Solutions Engineering Group lab, each functional team ran Cat 5 cables throughout the lab between 12-port patch blocks and centrally located patch panels. Next, AMIQ modules were attached to servers and connected to an AMX switch. When each server is powered up, it is automatically added to the KVM configuration database, which holds information about devices in the AMX switching matrix, such as changes made to system connections. This database is always stored locally on the AMX 50x0 switches, but a copy can also be stored on a separate server using Avocent AMWorks® software.

In some cases, AMIQ modules simply were attached to the console connector of an existing 8-port or 16-port analog KVM switch and added to the AMX switching matrix. For 16-port OutLook 2160ES switches that were already installed in a rack, functional teams continued using the rack-mounted LCD display on one 2160ES console port and connected the AMIQ module to the second port.

When a server is powered up with the new AMIQ module attached, the module automatically connects to the AMX switching matrix and appears in the Avocent On-Screen Configuration and Activity Reporting (OSCAR®) menu. Initially, the name provided for

the module in the OSCAR menu is the unique identifier (UID) number, which is located on a sticker on the back of each AMIQ module. Once the AMIQ module appears in the OSCAR menu, administrators can give it a more descriptive name.

For example, the name of the server to which the AMIQ module is attached—or the location of the server, such as “rack9-appsrv14”—can be set on the AMIQ module so that the server can be easily identified. The AMIQ module stores this setting in nonvolatile memory space, retaining the information even if the module is physically moved to a different system. By establishing a consistent naming convention, administrators can facilitate quick identification of the systems to which AMIQ modules connect.

To simplify initial system identification, administrators can power on one system at a time. The new AMIQ is easily locatable, because it will be the only module with a service profile identifier (SPID) number instead of an administrator-assigned name.

Enabling remote access using a Dell 2161DS switch

The 2161DS switch, which provides remote access to the AMX matrix, was installed and connected to the AMX switch through an AMX5100 console. That is, the server interface pod (SIP) module for the 2161DS switch was connected to the console port on the AMX5100, and the AMX5100 was connected to the output port on the AMX50x0 switch. Local administrators created usernames and passwords on the 2161DS switch, and remote administrators installed the Dell Remote Console Software (RCS) management application on their desktop or notebook systems. When connecting through a firewall, administrators must allow network ports 2068 for Secure Sockets Layer (SSL) authentication, 3211 for PS/2 devices, and 8192 for digital video.

MODELS IN THE CURRENT DELL AND AVOCENT SWITCH LINES

Dell and Avocent KVM switches are flash-upgradeable and include the OSCAR management tool, which provides a menu-based display from which administrators can select servers and system configurations. OSCAR supports cascading, so that switches can be used together in a matrix.

Avocent AMX switches enable multiple local administrators to view the same server simultaneously, although only one can control the server at any given time. Built-in support for the KVM over IP protocol allows the Dell 2161DS switch to support up to two simultaneous remote administrators in addition to one local console, although each administrator must access a different server.

Avocent AMX switches are especially useful in larger installations, such as the Dell Enterprise Solutions Engineering Group lab, because they have a large number of server ports (inputs) and consoles (outputs) and support matrix extension through connection to additional AMX switches. The Avocent AMX switches provide real-time video output up to 1,000 feet from the servers.

Dell 2161DS Remote Console Switch

This 1U rack-mountable switch supports one local console (output) and 16 server ports (inputs). Using the optional eight-port port expansion module (PEM), administrators can expand each of the 16 server ports on the 2161DS switch to eight additional server ports. By connecting as many as eight servers to a PEM, administrators can support up to 128 servers using one 2161DS switch.

The 2161DS switch reduces cable bulk in rack installations by using standard Cat 5, Cat 5e, or Cat 6 UTP cabling in conjunction with server interface pods (SIPs), which connect the switch to individual systems. The SIPs condition and amplify the analog signal before sending it through a single Cat 5, Cat 5e, or Cat 6 cable. The 2161DS switch supports remote access through a 10/100BaseT Ethernet port. Remote administrators run the Dell Remote Console Software (RCS) Java-based client interface to initiate a digital KVM over IP session through the 2161DS switch. A maximum video resolution of 1280 × 1024 is supported for remote connections.

Avocent AMX5000 switch

This 1U rack-mountable switch supports up to eight local consoles (outputs) through the AMX5100 interface (see "Avocent AMX5100 user station"). Each console handles up to 32 systems (inputs). Up to eight administrators can simultaneously access any of the 32 systems from any of the AMX5100 user stations, or they can access the same system. As a single unit, the AMX5000 switch offers a non-blocking configuration. The 32 inputs can be used to connect to the outputs of additional AMX5000 or AMX5010 switches (see "Avocent AMX5010 switch") to support extra systems in a variety of blocking and nonblocking configurations. AMX5000 and AMX5010 switches also can be cascaded to Avocent OutLook 180ES or OutLook 2160ES switches by connecting an Avocent AMIQ module—which is similar to a SIP—to a 180ES or 2160ES output port.


Avocent AMX5010 switch

This 2U rack-mountable switch supports up to 16 local consoles (outputs) through the AMX5100 interface. Each console handles up to 64 systems (inputs). Up to 16 administrators can simultaneously access any of the 64 systems from any of the AMX5100 user stations, or they can access the same system. Similar to the AMX5000 switch, the AMX5010 can connect to additional switches to support extra systems in various blocking and nonblocking configurations.

Avocent AMX5100 user station

The AMX5100 interface for the local consoles is usually placed on the administrator's desk and connects to the local monitor (VGA port), PS/2 keyboard, and PS/2 mouse port. The AMX5100 user station also connects to one of the outputs of the AMX5000 or AMX5010 switch. To enable remote IP access to AMX switches for one user, administrators also can connect an AMX5100 console to a KVM over IP switch product such as the Dell 2161DS switch or the Avocent DS1800 switch, which supports eight local consoles (outputs).

Enabling efficient hardware sharing and secure remote access

Dell and Avocent switching products offer the flexibility to implement analog and digital KVM connections in enterprise data centers. By implementing the Dell 2161DS Remote Console Switch and Avocent AMX switching products in the Dell Enterprise Solutions Engineering Group lab, Dell engineers gained an efficient way to access and share equipment in a matrix-switching environment—and enabled secure, remote logins that allow them to control systems when they are away from the lab. 

Mike Kosacek (michael_kosacek@dell.com) is a senior member of the Custom Solutions Engineering Group at Dell. He specializes in storage area networks (SANs) and clustering. Previously, Mike has been the lead development engineer on several Dell high-availability cluster products. He has an Electronics Technology degree and is a Microsoft Certified Systems Engineer (MCSE).

Installing AMWorks software on Dell servers for backup and administration. An AMWorks software license is included with each AMX5000. This Java™-based system administration tool supports customized user profiles and multilevel security, and allows system administrators to build a user database with assigned access and password protection for each server. The user database holds the user logins and access controls.

To ensure that the configuration could be easily backed up and to simplify any future administration, the user and configuration databases that were created on the AMX switches were synchronized to a local database on a Dell server that had AMWorks software installed. This copy serves as a backup, in case a failed AMX switch needs to be replaced, and also allows new administrators and devices to be added and the new configuration pushed out, or *synchronized*, to the AMX switches. AMWorks software also allows firmware upgrades to be installed remotely, without the need to physically attend to each switch or AMIQ module.

Integrating Nagios and Ganglia

with Dell OpenManage Server Administrator in Linux-based Environments

Enterprises that run the Red Hat® Linux® operating system on Dell™ PowerEdge™ servers can monitor system health proactively using open source tools such as Nagios and Ganglia. This article explains how to integrate the monitoring capabilities of Nagios and Ganglia with the Dell OpenManage™ Server Administrator command-line interface.

BY DAN BERES, ROGER GOFF, AND TERRY SCHROEDER

Today's data centers and high-performance computing (HPC) clusters pack more computing power into less physical space. The increased density can generate a significant amount of additional heat that often is not accompanied by increased cooling capacity. Current-generation Intel® processor-based systems react to higher operating temperatures by increasing fan speeds and, if temperatures get hot enough, limiting CPU performance to reduce power consumption. In the case of HPC clusters running parallel applications, a reduction in CPU performance of just one system may slow the performance of the entire cluster. By monitoring system health proactively, administrators can detect and address problems before they affect application performance.

Administrators typically assess the current state of network devices in one of two ways. The *pull* method queries device instrumentation from a central monitoring console at specified intervals, receiving a status or explicit data value in response. Alternatively, the *push* method constantly reports on device status by sending Simple Network Management Protocol (SNMP) traps, data, or

both from device instrumentation to a central monitoring console. Both pull and push methods offer advantages.

Dell™ OpenManage™ Server Administrator provides instrumentation for Dell PowerEdge™ servers and monitors systems using the push method. This article focuses on Nagios and Ganglia, two open source monitoring tools that use the pull method and can integrate with Dell OpenManage Server Administrator to help manage PowerEdge servers. Nagios—a host, service, and network monitoring program designed to quickly inform system administrators of problems—uses Linux® shell scripts and executables to retrieve and report state information. Ganglia is a real-time, agent-based monitoring tool for HPC systems.

Managing servers with the Dell OpenManage Server Administrator CLI

Dell OpenManage Server Administrator provides multiple interfaces and integrates with open source frameworks such as Ganglia and Nagios. The most common interfaces are SNMP and the command-line interface (CLI). The examples shown in this article—for Nagios, a custom

```
> omreport chassis

Health

Main System Chassis

SEVERITY      : COMPONENT
Ok            : Fans
Ok            : Intrusion
Ok            : Memory
Ok            : Power Supplies
Ok            : Temperatures
Ok            : Voltages
Ok            : Hardware Log

>
```

Figure 1. Example report from omreport chassis command

plug-in that monitors the thermal status of any Dell PowerEdge server; for Ganglia, a collection script that gathers and records temperatures for all temperature probes in a PowerEdge server—can be leveraged to create plug-ins and scripts that monitor other system health components of a PowerEdge server. To run commands from the CLI, Dell OpenManage Server Administrator must be installed on each server being monitored.¹

System administrators can obtain a quick system health overview of a PowerEdge server by running the CLI command `omreport chassis`. Figure 1 shows the type of information this command generates. When a problem occurs in a server component, the report indicates a severity other than “OK” next to the component name. To procure more in-depth information, administrators can add additional arguments to the `omreport chassis` command. For example, Figure 2 displays an excerpt from the main system chassis temperature report for a PowerEdge 2650 server, which was generated by the `omreport chassis temps` command. This sample displays readings from only the first temperature sensor.

Although the format for the `omreport chassis temps` report is the same regardless of which PowerEdge server it runs on, the output for each PowerEdge server model differs by the number and names of temperature probes reported. This article presents shell scripts that address the variability in probe numbers and names, enabling administrators to deploy both the Nagios and Ganglia examples across all PowerEdge servers.

Figure 3 contains a listing of other PowerEdge attributes that administrators can collect using Dell OpenManage Server Administrator.

```
> omreport chassis temps

Temperature Probes Information

-----
Main System Chassis Temperatures: Non-Critical
-----

Index          : 0
Status         : Ok
Probe Name     : ESM Frt I/O Temp
Reading        : 26.0 C
Minimum Warning Threshold : 10.0 C
Maximum Warning Threshold : 50.0 C
Minimum Failure Threshold : 5.0 C
Maximum Failure Threshold : 55.0 C

:
:
```

Figure 2. Excerpt from omreport chassis temps command report

```
> omreport chassis -?

Command      Description
acswitch     AC switch settings.
bios         BIOS properties.
biossetup    BIOS setup configuration.
currents     Current probe(s) properties.
fans         Fan probe(s) properties.
fancontrol   Fan control settings.
leds         Chassis LED settings.
firmware     Firmware properties.
info         Chassis information.
intrusion    Chassis intrusion status.
memory       System memory configuration.
nics         Network interface card(s) properties.
ports        Port(s) properties.
powerbutton  Power button control settings.
processors   Processor(s) properties.
pwrsupplies  Power supply(s) properties.
slots        Slot(s) properties.
temps        Temperature probe(s) properties.
volts        Voltage probe(s) properties.

>
```

Figure 3. Example report from omreport chassis -? command

¹ For a detailed listing of what information can be reported through the Dell OpenManage Server Administrator CLI, visit the *Dell OpenManage Server Administrator Command-Line Interface User's Guide* online at <http://docs.us.dell.com/docs/software/svradmin>.

This report was generated by running the `omreport chassis -? command`. The remainder of this article focuses on data that is returned when executing the `omreport chassis temps` command.

Monitoring system health with Nagios

Nagios is a monitoring console with a Web interface that can display system health in a one-to-many relationship. It is available under the GNU General Public License (GPL) from <http://www.nagios.org>. Although Nagios can receive SNMP information, its true strength is the ability to use any application or script to gather data. Data gathering programs, often called plug-ins, are placed and run on each client system being monitored. Plug-ins return relevant data and one of three states: "OK," "Warning," or "Critical." Administrators schedule plug-ins to run at specified polling intervals.

For the primary logic of a plug-in that returns the thermal status of a PowerEdge server, visit *Dell Power Solutions* online at http://www.dell.com/magazines_extras. To determine the thermal state and report it to Nagios, the plug-in compares actual system temperatures to the temperature thresholds set in the BIOS for each thermal sensor within the server. The bash shell script—which reports PowerEdge server thermal status and manages input parameter parsing, usage statements, and so forth—is called `nagios_check_temps` and can be found online at http://www.dell.com/magazines_extras.

Nagios takes the status returned from a plug-in, displays it on the Web console, and stores its value in a database, enabling administrators to query the database for system status over specified time intervals. Nagios also can trigger alert actions—from sending an e-mail message to launching a script—based on the reported status. Figure 4 shows the Nagios Web console, which displays the output of the `check_temps` plug-in run on two PowerEdge servers.

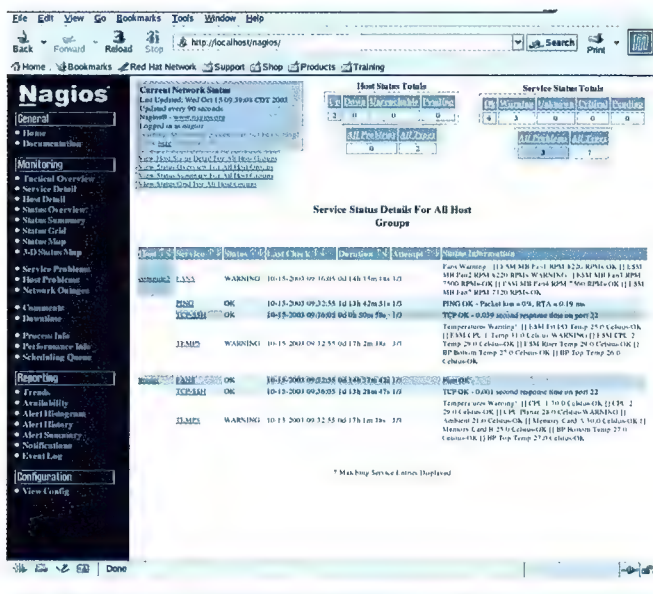


Figure 4. Nagios Web console showing display for the `check_temps` plug-in

Dell OpenManage Server Administrator provides multiple interfaces and integrates with open source frameworks such as Ganglia and Nagios.

Nagios is an excellent tool for alert and trend management of PowerEdge servers. Its ability to gather any information, and subsequently to report on the status of that information, is particularly valuable for organizations seeking to integrate the management of their IT infrastructure. This article presents only a rudimentary example of what Nagios can do with the information gathered from the Dell OpenManage Server Administrator utilities. Many other applications are viable, making Nagios an effective tool for managing PowerEdge servers.

Monitoring cluster metrics with Ganglia

The open source forum SourceForge maintains Ganglia, a widely accepted cluster-monitoring program. Ganglia provides a Web-based front end to display real-time data for both an aggregate cluster and each system in a cluster. A multithreaded daemon process runs on each cluster node to collect and communicate the host state in real time.

By default, Ganglia monitors a collection of metrics, including CPU load, memory usage, and network traffic. It also provides a tool called `gmetric` that enables administrators to extend the set of metrics they monitor. To monitor sensors on PowerEdge servers, Ganglia pulls values from each server using the Dell OpenManage Server Administrator CLI and passes those values to `gmetric`. Administrators can track the values from the Ganglia Web console. This article discusses Ganglia version 2.5, but a new version of Ganglia currently under development will modify the method for extending the data that Ganglia monitors.

The following command line uses the Dell OpenManage Server Administrator `omreport` statement, combined with the Linux `grep` and `awk` commands, to retrieve the temperature of CPU 1 on a PowerEdge 2650 server:

```
omreport chassis temps index=1|grep Reading|awk
' { print $3 }'
```

If administrators create a command called `cpu1_temp` that returns the output of the previous `omreport` command line, then the following command line will send the result to the Ganglia Web console and create a new metric graph labeled "cpu1_temp":


```
gmetric -name cpu1_temp -value 'cpu1_temp'
-type float -units Celsius
```

Running gmetric once inserts a single result into the Ganglia database and plots a single data point on the metric graph in the Ganglia Web interface. To track a sensor's value over time, administrators must place entries into cron² to execute gmetric at the desired polling interval.

Administrators can track all the temperatures on a PowerEdge server using a Linux bash shell script. For the bash shell script, `ganglia_check_temps`, visit http://www.dell.com/magazines_extras. Figure 5 shows the resulting Ganglia display after setting up system cron jobs to run this temperature collection script on a PowerEdge 2650. Similar scripts to monitor system health parameters, such as fan speeds and voltages, can be derived from this script.

Ganglia can be a powerful tool for cluster administrators who need to track the resource utilization and system health of cluster nodes. The data trends provide administrators with valuable information that can be used to identify issues and plan future system and data center requirements.

Protecting server health with powerful tools

Administrators in enterprise Linux environments rely heavily on tools such as Nagios and Ganglia to proactively monitor the state and utilization of compute resources. When combined with the Dell OpenManage Server Administrator CLI, Nagios and Ganglia enable early detection of device conditions that may be an indicator of broader data center problems. Early detection can help system administrators proactively address warning conditions before they lead to system failure and unplanned downtime. 

Dan Beres (daniel_beres@dell.com) is an enterprise technologist in the Advanced Systems Group at Dell. His areas of interest include programming in C and Assembler, building management controls of complex systems, and integrating Linux into the enterprise. He has been in the computer industry for more than 20 years, spending the past five and a half years at Dell. Dan has a B.F.A. in Communications from Chapman College.

Roger Goff (roger_goff@dell.com) is an enterprise technologist in the Advanced Systems Group at Dell. His current interests include Linux and Microsoft® Windows® HPC clusters and cluster file systems. Roger is a Red Hat® Certified Engineer and has an M.S. and a B.S. in Computer Science from Virginia Polytechnic Institute and State University (Virginia Tech).



Figure 5. Example Ganglia displays after running temperature collection script

Terry Schroeder (terry_schroeder@dell.com) is an enterprise technologist in the Advanced Systems Group at Dell. He supports Dell field system consultants and engineers by communicating Dell systems management products and initiatives to customers. Terry came to Dell with nine years of corporate IT experience, having held numerous management and implementation positions covering infrastructure and application initiatives. Terry has an M.S. in Library Science and Information Management and a B.S. in Social Sciences, both from Emporia State University.

FOR MORE INFORMATION

Nagios:
<http://www.nagios.org>

Ganglia:
<http://ganglia.sourceforge.net>

² Cron is a program that allows administrators to create jobs that will run at a given time.

Remotely Managing UNIX and Linux Servers Using the Dell RAC Serial/Telnet Console

Introduced in the 3.0 release of Dell™ remote access controller (RAC) firmware, the RAC serial/telnet console provides administrators with a standard serial console for remotely managing Dell PowerEdge™ servers running UNIX® or Linux® operating systems. The console offers system power management capabilities, pre-operating system redirection capabilities, and support for kernel-mode messaging.

BY AURELIAN DUMITRU

Dell™ remote access controllers (RACs) provide remote systems management capabilities on supported Dell PowerEdge™ servers. Although four types of RACs are available—Dell Remote Access Card III (DRAC III), DRAC III/XT, Embedded Remote Access (ERA), and Embedded Remote Access Option (ERA/O)—they share many of the same features and are supported by a common software stack, which is part of Dell OpenManage™ Server Administrator.

One new feature introduced in the 3.0 release of RAC firmware is the RAC serial/telnet console, which provides administrators with a standard serial console for remote management of servers running UNIX® or Linux® operating systems. Dell has enhanced its standard serial console with new functionality, none of which requires custom software to be installed or running on the host:

- System power management to enable power up, power down, power cycle, or reset
- Redirection of pre-operating system firmware such as BIOS screens
- Support for kernel-mode messaging using SysRq magic keys

This article explains how the RAC serial/telnet console is implemented, provides details on how to configure the console, and includes an example of how to set up the IT environment for optimal console performance.

Understanding the functionality of the RAC serial/telnet console

The RAC serial/telnet console provides a means by which administrators can access the ttyS1 serial console and video console of a UNIX or Linux host through either a serial connection (using a VT-100 or ANSI® client) or a LAN-based Telnet connection. The RAC serial/telnet console provides a rich set of commands for performing system power management tasks, configuring the RAC, and viewing RAC and system logs. The serial/telnet console uses the same set of commands as the Racadm command-line utility that originally was part of the RAC software stack. Therefore, administrators need not learn a new set of commands to use the RAC console.

The RAC acts as a gate between its PowerEdge host system and the remote administrator, providing optimal console-redirection performance combined with user-level security. Depending on the hardware interfaces and support offered by its host, the RAC can provide

access to the host console through a serial connection to the RAC, a Telnet connection, or both. For example, provided that the RAC is properly configured, the Dell PowerEdge 2650 server offers connectivity using either a VT-100 or ANSI serial client or a LAN-based telnet client.

As shown in Figure 1, the RAC hardware features two serial interfaces (RAC serial 1 and RAC serial 2) and one video interface (RAC video interface). The RAC serial 1 interface is used for host-to-RAC serial communication; the RAC serial 2 interface is used for RAC-to-client communication. Thus, administrators must properly set up two baud rates (baud_rate_1 and baud_rate_2) to gain remote access to the serial console on the host. Baud_rate_1 in the RAC configuration must be in sync with the Linux ttyS1 baud rate. Baud_rate_2 must be in sync with the baud rate of the VT-100 or ANSI client. Any mismatch between these values will make the RAC serial/telnet console unusable. Note that baud_rate_1 and baud_rate_2 are not necessarily set to the same value.

For optimal performance, Dell recommends setting baud_rate_1 to 57,600 bps and baud_rate_2 to 115,200 bps, as well as enabling hardware flow control for the host. RAC relies on hardware flow control to avoid dropped characters during serial communication; flow control enables the host to respond to Request To Send/Clear To Send (RTS/CTS) signals.

Regardless of whether a serial or Telnet connection is established with the RAC, the console will present the user with a login prompt. The login is authenticated against the RAC user database, which resides on the RAC. Because the authentication is RAC-based, not host-based, administrators can log in to the RAC console even on a dead server, provided that the server has power. Once authenticated, the administrator connects to the serial console using the connect com2 command, or to the video console of the host using the connect video command.

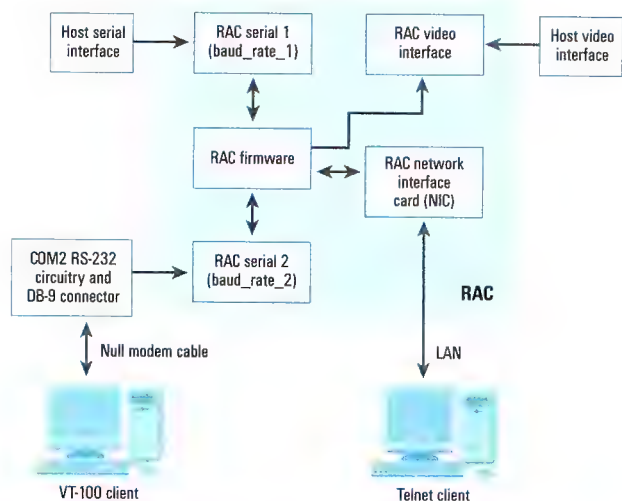


Figure 1. RAC serial/telnet console configuration on PowerEdge 2650 server

To maintain security, all users are not granted all privileges. All administrators should be able to access the serial and video consoles of the host. However, only the root user can power up the system or configure the RAC—for instance, to change a baud rate. In addition, the root user's password and location within the RAC user database can be changed.

Configuring the RAC serial/telnet console

To set up the RAC serial/telnet console, the following RAC configuration parameters need to be modified in the following sequence:

1. Upgrade the RAC firmware to version 3.0.
2. Upgrade the host system BIOS to the latest version available, or check to make sure that the current system BIOS supports the RAC serial/telnet console.
3. Configure and enable the RAC serial/telnet console by saving the following sample RAC configuration in a text file (for instance, rac.cfg) and then applying the configuration using the Racadm interface:

```
#
# Object Group "cfgSerial"
#
[cfgSerial]
cfgSerialBaudRate=115200
cfgSerialConsoleEnable=1
cfgSerialConsoleQuitKey=.,/
cfgSerialTelnetEnable=1

#
# Object Group "cfgRacTuning"
#
[cfgRacTuning]
cfgRacTuneHostCom2BaudRate=57600
```

The preceding sample configuration file will:

- Enable the RAC serial console
- Enable the RAC telnet console
- Set the RAC-to-client baud rate to 115,200 bps
- Set the host-to-RAC baud rate to 57,600 bps
- Set the console quit key sequence—which allows the administrator to exit the host ttyS1 serial console or video console redirection and go back to the RAC serial/telnet console—to “.,/” (comma dot slash)

In the sample configuration file, the names in square brackets are the groups into which the various configuration parameters


```

serial --unit=1 --speed=57600
terminal --timeout=10 console serial

title Red Hat Linux Advanced Server (2.4.9-e.3smp)
    root (hd0,0)
    kernel /boot/vmlinuz-2.4.9-e.3smp ro root=/dev/sda1 hda=ide-scsi console=tty0 console=ttyS1,57600
    initrd /boot/initrd-2.4.9-e.3smp.img
title Red Hat Linux Advanced Server-up (2.4.9-e.3)
    root (hd0,0)
    kernel /boot/vmlinuz-2.4.9-e.3 ro root=/dev/sda1 s
    initrd /boot/initrd-2.4.9-e.3.im

```

Figure 2. Modifications to /etc/grub.conf to enable ttyS1 serial console redirection

are organized. Refer to the *Dell Remote Access Controller Racadm User's Guide* for additional information on RAC configuration parameters.

4. Use the following `racadm` command to apply these changes:

```
racadm remote connect options config -f
    config filename
```

where `remote connect options` is `-r rac_ip_addr -u root -p password` and `config filename` is the name of the configuration file (for instance, `rac.cfg`). The `remote connect options` variable is needed only for executing Racadm remotely, for example when configuring a remote RAC from a management station over a LAN. From the RAC's host console, administrators need not use this variable because Racadm can communicate directly with the RAC that exists in that system.

5. Reset the RAC using the following command to make the changes effective:

```
racadm remote connect options racreset
```

6. Reboot the RAC host, enter the BIOS configuration page, and select "Integrated Devices." On the PowerEdge 2650, verify that Serial Port 2 is set to COM2. On the PowerEdge 1750, verify that Serial Port 1 is set to "Off."

7. Select "Console Redirection" from the main BIOS configuration page. The three settings should be configured as follows for the PowerEdge 2650 server:

- Console Redirection: Serial Port 2
- Remote Terminal Type: VT-100
- Redirection After Boot: Enabled

For the PowerEdge 1750 server, the administrator must choose RAC instead of Serial Port 2.

8. Save the changes and reboot the system. The system reboot is important because it allows the RAC and the BIOS to synchronize their communication parameters. Depending on the specific configuration of a system, the RAC may not redirect BIOS screens if the BIOS and the RAC do not sync up. Because the BIOS and the RAC synchronize when the system boots, at least one system boot needs to occur after the administrator enables and configures the RAC serial/telnet console. No additional reboots are required thereafter.

9. After both the RAC and the BIOS are in sync, at least three Linux files must be changed to enable ttyS1 serial console redirection: `/etc/grub.conf` or `/etc/lilo.conf` (depending on the bootloader used), `/etc/inittab`, and `/etc/securetty`. However, none of these changes is required if only access to the host video console is needed. Required modifications to the three files are shaded in Figures 2, 3, and 4.

Note: In Figure 2, the Linux host baud rate is set to 57,600 bps in the `/etc/grub.conf` or `/etc/lilo.conf` file, which is in sync with the host-to-RAC baud rate set in the sample RAC configuration text file (`rac.cfg`) in the

```

# Run gettys in standard runlevels
co:2345:respawn:/sbin/agetty -h -L 57600 ttyS1 vt100
1:2345:respawn:/sbin/mingetty tty1
2:2345:respawn:/sbin/mingetty tty2
3:2345:respawn:/sbin/mingetty tty3
4:2345:respawn:/sbin/mingetty tty4
5:2345:respawn:/sbin/mingetty tty5
6:2345:respawn:/sbin/mingetty tty6

```

Figure 3. Modifications to /etc/inittab to enable ttyS1 serial console redirection

```

.
.
.
tty8
tty9
tty10
tty11
ttyS1

```

Figure 4. Modifications to `/etc/securetty` to enable `ttyS1` serial console redirection

previous example. The same host-to-RAC baud rate of 57,600 bps is used when invoking the `agetty` command (see Figure 3). Nevertheless, the `-h` flag, when present on the same line, forces `agetty` to use hardware flow control for the communication. As explained in “Understanding the functionality of the RAC serial/telnet console,” the RAC relies on hardware flow control for reliable serial communication with its host.

10. After modifying the files, restart `agetty` or reboot the RAC host to complete the configuration process. The RAC is now ready to redirect its host serial console.

Optimizing the IT environment for serial/telnet console performance

Depending on the IT environment and the hardware capabilities of the RAC host, administrators may choose to use either a serial connection or a Telnet connection to the RAC. Many organizations use serial concentrators, or *terminal servers*, to consolidate multiple serial connections. Unfortunately, not all serial concentrators are alike: they vary with respect to configuration parameters, configuration interface, and wiring. Most serial concentrators use an RJ-45-to-DB-9 adapter. For connection schematics for two adapters that can be used to connect a RAC to a Cisco® 2511 serial concentrator, visit *Dell Power Solutions* online at http://www.dell.com/magazines_extras.

Regardless of whether a serial or Telnet connection is used, administrators must first log in to the RAC prompt to view RAC system logs, then redirect either `ttyS1` serial consoles or video consoles, and—if logged in as root—configure the RAC or reboot the system. Some organizations also use SysRq magic keys to capture kernel-level status information as part of the IT support process.

The RAC serial/telnet console supports SysRq functionality, although the methodology is slightly different for a remote connection than for a local console. At the local console, the administrator normally would press the `Alt + SysRq + magic_key` sequence to get the expected SysRq output. If using the remote connection, the

administrator enters the `send break` command on the serial or telnet client and then presses one of the SysRq magic keys. This SysRq functionality will work only when redirecting `ttyS1`; it will not be present if the administrator connects to the remote console and then redirects the host's video console. To gain access to `ttyS1`, administrators must use the `connect com2` command in the RAC console.

For example, an administrator may want to collect as much information as possible to assist with a problem diagnosis and to store this information in a file. Generally the problems encountered can be divided into two areas: an unresponsive system or a system that has crashed. In either case, the administrator may need to connect to the RAC and then access the host `ttyS1` serial console or video console. If the system appears to be hung, the administrator probably requires output from the “b,” “m,” “p,” and “t” SysRq magic keys. Once that information is gathered, the RAC can be used remotely to restart the server, after which the administrator can examine the boot screens for further clues. If the system appears to have crashed, then the best approach may be to examine the RAC system logs. Depending on the state of the system, the administrator may choose to use the RAC to restart the server.

Enabling server recovery and configuration management through RAC

The RAC serial/telnet console can be a valuable remote management tool for IT administrators. Besides supporting industry-wide systems management practices, the RAC serial/telnet console is enhanced with features that allow for server recovery, system configuration management, and RAC configuration management. The RAC serial/telnet console is software-independent and helps administrators interact smoothly with the RAC host. ☺

Aurelian Dumitru (aurelian_dumitru@dell.com) is a senior software engineer with the Custom Solutions Engineering team at Dell, where he works to deploy and customize systems management for medium to large enterprises. Before joining the Custom Solutions Engineering team, he held the lead engineer position for the Remote Management Delivery team. Aurelian has 12 years of experience in hardware, software, and system design and integration. He has an M.S.E.E. degree from the Technical University of Iasi, Romania.

FOR MORE INFORMATION

Dell Remote Access Controller Racadm User's Guide:

<http://docs.us.dell.com/docs/software/smdrac3/RAC/en/index.htm>

Remote serial console how-to:

<http://www.faqs.org/docs/Linux-HOWTO/Remote-Serial-Console-HOWTO.html>

Troubleshooting Servers

with Dell Remote Access Controllers

System downtime is a serious detriment to business—enterprises stand to lose money and customers if a server fails and business-critical applications are interrupted. Using Dell™ remote access controllers, system administrators can identify server problems quickly and resolve them efficiently.

BY JON MCGARY

Server failure is a reality that all system administrators must face. However, armed with the right tools and the knowledge of how to use them effectively, administrators can identify and resolve common problems quickly. Dell™ remote access controllers (RACs) provide alert notification; event traceability through the use of log files; and remote server access using power management, console redirection, and remote floppy boot tools to assist administrators in troubleshooting and correcting problems.

Four types of RACs are available: Dell Remote Access Card III (DRAC III), DRAC III/XT, Embedded Remote Access (ERA), and Embedded Remote Access Option (ERA/O). Each RAC is designed to provide extensive remote management capabilities for Dell PowerEdge™ servers.

Alerting administrators to potential issues

By notifying administrators immediately when a warning condition arises, a RAC can help reduce and sometimes avoid system downtime. The RAC continuously monitors server sensors and their Intelligent Platform Management Interface (IPMI) hardware logs to determine when to send alerts to administrators. IPMI—a specification for management controllers that are embedded in system components—enables RACs to monitor the physical condition of servers based on sensor reports of characteristics such as temperature, voltage, fan, power supply, and cover intrusion.

Administrators can configure the RAC to send an alert to as many as 16 recipients when a monitored component exceeds a specified operating range; alerts can be sent through the RAC's integrated network adapter or modem (the latter is supported only by DRAC III). Supported alert notification formats are e-mail messages, Simple Network Management Protocol (SNMP) traps, and alphanumeric and numeric pages (see Figure 1).

Each alert will be classified at one of three possible status levels: informational (healthy), warning (noncritical), or critical (failure). An informational alert indicates that the system is healthy or working as expected. This type of alert may be generated when, for example, a system moves from an unacceptable operating temperature (warning or critical range) back to an acceptable temperature. A warning alert indicates that a monitored component is outside an administrator-specified operating range. A critical alert indicates that a component is in a failure condition—it has moved outside the minimum or maximum threshold values and requires immediate attention.

An e-mail alert contains the following information: message (including test message), event description, date, time, severity, system ID, model, BIOS version, asset tag, service tag, managed system name, operating system (OS) type and name, and Embedded Systems Management (ESM) version.

RAC type	E-mail message	SNMP trap	Alphanumeric page	Numeric page
DRAC III with modem	Yes	Yes	Yes	Yes
DRAC III/XT	Yes	Yes	No	No
ERA	Yes	Yes	No	No
ERA/O	Yes	Yes	No	No

Figure 1. Supported formats for alert notification

Accessing a server through the RAC

Once administrators are alerted to potential problems, they can connect remotely to the RAC to review the trace logs. RACs provide the following methods to access a failed server or a server that cannot be accessed through its network adapter; this article focuses on the features of the RAC's Web-based interface.

- **Web-based interface:** Enables remote graphical access through the RAC's network adapter using a supported browser.
- **Racadm command-line interface (CLI) utility:** Allows administrators to connect to the managed server and execute Racadm subcommands from a remote console or management station using only the IP address of the managed server. This utility is supported only on Microsoft® Windows Server™ 2003, Microsoft Windows® 2000 Server, and Red Hat® Linux® operating systems.
- **Telnet console:** Provides access through the RAC network adapter to the COM2 port, video, and hardware management interfaces of the server, and supports serial and Racadm commands to the RAC including system boot, reset, power up, and power down.
- **Terminal emulation:** Provides access through the RAC to the COM2 port, video, and hardware management interfaces of the server through either the external serial connector or external DRAC III VT-100 serial connector. Terminal emulation software such as Hilgraeve HyperTerminal® or minicom supports serial and Racadm CLI utilities.

Connection to the RAC through the Web-based interface

Administrators can use any supported Web browser to access the RAC's Web-based interface—no additional software is required. To connect to the login window, administrators enter "http://RAC IP address" (where RAC IP address is the IP address of the RAC) in the address field of the Web browser. The RAC login window will appear; typing the local RAC username and password will authenticate and log the administrator into the RAC.

Investigating the problem

Once the RAC has identified a potential problem and sent a notification, system administrators can investigate the alert condition using the hardware log, RAC log, power-on self-test (POST) log, and last-crash screen.

Hardware log. The hardware log, also referred to as the ESM log, displays critical events that occur on the managed server, such as component threshold changes, system reset, and system boot (see Figure 2). The log is the first place an administrator should check if the system is not functioning properly. The hardware log is generated by ESM instrumentation on the server and by the RAC if it is configured to monitor any managed system events.

RAC log. The RAC log is a persistent log maintained in the RAC firmware. It contains a list of user actions and alerts issued by the RAC (see Figure 3). If the RAC loses communication with the managed server, all entries that the RAC would have added to the hardware log (such as power failure or RAC sensor alerts) are added to the RAC log until communication is reestablished.

POST log. The POST log is helpful when troubleshooting an abnormal boot process. It records the events and processes that occur when the server boots but before the OS starts. Activities such as configuring SCSI and RAID controllers and initializing Peripheral Component Interconnect (PCI®) cards are some of the tasks that occur before the OS starts. The contents of the POST log are written by the BIOS of the managed server and are overwritten during each system boot. The log displays a POST code and description of the boot event.

Last-crash screen. The RAC last-crash screen feature can be configured to capture the last image on a managed Windows-based server when a system hangs or crashes. This display provides information about events that led to the system crash. If the OS stops communicating with the RAC, a snapshot image of the

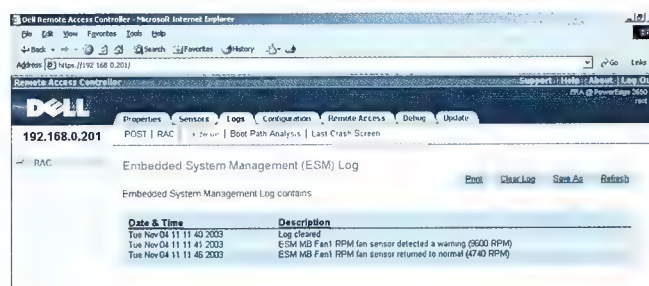


Figure 2. The hardware log

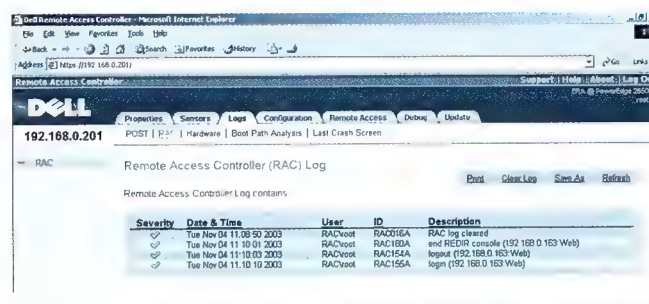


Figure 3. The RAC log

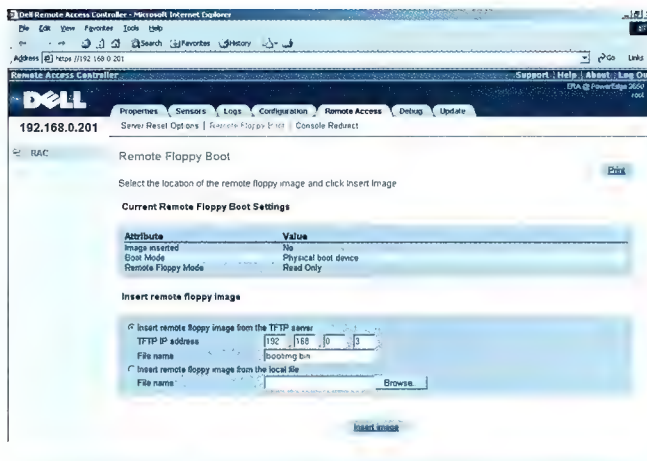


Figure 4. The Remote Floppy Boot screen

server console is retrieved by the RAC and stored in persistent RAC memory for review.

Remotely managing the server using the RAC

Server reset options, console redirection, and the RAC remote floppy boot (RFB) feature can help administrators resolve problems remotely once alert conditions have been identified.

Server reset options

In some situations, the administrator must reboot the server to resolve the problem or to monitor the boot process (using text console redirection) based on information found in the POST log from the previous system boot. The RAC provides several power management actions on the managed server, such as graceful shutdown through the OS or hard reset. From the Server Reset Options window in the RAC Web-based interface, the following power management options are available:

- **Graceful server shutdown:** Shuts down the managed server using OS interfaces. If the OS is not available, this option is grayed out on-screen and unavailable.
- **Graceful server restart:** Shuts down and restarts the managed server using OS interfaces. If the OS is not available, this option is grayed out on-screen and unavailable.
- **Reset:** Resets the system (equivalent to pressing the reset button); the power is not turned off by this action.
- **Server power cycle:** Powers down the server and powers it up again (equivalent to pressing the power button twice).
- **Server power on:** Powers up the server (equivalent to pressing the power button once).
- **Server power off:** Powers down the server (equivalent to pressing the power button once).

Console redirection

After connecting to the RAC, administrators can access the text and graphical consoles¹ of the managed server using the Console Redirect tab from the Remote Access window of the RAC Web-based interface. The console redirection option allows administrators to use the display, mouse, and keyboard on a local management station to control the corresponding devices on the remote managed server. This feature is critical for administrators because it allows them to open a text redirection window on the remote server to control and monitor the boot sequence, helping identify initialization failures during the boot process.

Remote floppy boot

If additional system diagnostics or tools must be loaded on the server to assist in diagnosing problems, administrators can use the RFB feature of the RAC (see Figure 4). The RFB feature enables administrators to load a DOS-bootable floppy image into the RAC memory, reboot the server (using the RAC power management features), and restart the server from the floppy image. The RFB process is equivalent to inserting a DOS-bootable floppy disk into the remote server and booting from the disk.

The RFB feature allows administrators to create several bootable image files with various diagnostics or tools and to load the image that is most applicable to solve the problem. The RAC RFB window provides an interface for administrators to view current RFB settings as well as to insert, configure, and remove a bootable floppy image in the RAC.

Achieving new efficiencies using Dell RACs

Dell remote access controllers can help improve systems availability by alerting administrators when potential problem conditions are detected. In addition, RACs can log characteristics of critical system components and provide remote access to servers to help administrators troubleshoot and correct problems. As business-critical applications demand more system resources and higher availability, RACs are becoming indispensable to the system administrator's toolkit. Dell RACs can help resolve server failures effectively, reducing downtime and enhancing overall system efficiency. ☁

Jon McGary (jon_mcgary@dell.com) is a senior software developer in Dell OpenManage™ Remote Management. Prior to joining Dell, Jon was employed by Tandem Computers and specialized in remote management of fault-tolerant computers. He has a B.S. from Texas A&M University.

¹ If the Novell® NetWare® OS is installed on the managed server, only text-mode redirection is supported.

Simplifying Enterprise Deployment

of Dell Remote Access Controllers

The Racadm command-line utility can improve IT efficiency in enterprises that have large deployments of Dell™ PowerEdge™ servers by enabling administrators to configure and replicate settings across multiple Dell remote access controllers (RACs). Examples of activities that can be automated using Racadm include managing usernames and passwords, configuring RAC event management, and updating RAC firmware for a set of RACs.

BY ZAIN KAZIM, BALA BEDDHANNAN, AND ALAN DAUGHETEE

Dell™ remote access controllers (RACs) provide secure access to help improve local or remote systems management of Dell PowerEdge™ servers, whether operational or not. In enterprise environments that have large-scale PowerEdge server deployments, ongoing operations and updates to multiple RACs can be time-consuming. Powerful RAC management tools can improve IT productivity and resource management. Dell offers several options to simplify the management and maintenance of RACs. This article discusses the implementation of automated processes to manage, configure, and update multiple RACs using the Racadm command-line utility and native operating system (OS) scripts.

The Racadm command-line utility provides a scriptable interface that allows administrators to configure RAC settings either locally at a server (a *managed system*) or remotely from a console (a *management station*). The Racadm utility enables administrators to gather server status information, issue server control commands, and manage firmware stored in the RAC. Because Racadm runs on most popular operating systems, the utility provides consistent commands for a wide variety of platforms and architectures.

The Racadm utility supports operations through the use of command-line parameters, switches, and a configuration file that stores all pertinent data required to configure a RAC (see “Sample RAC configuration file parameters”). Using Racadm subcommands and the RAC configuration file, system administrators can configure and replicate settings across multiple RACs with the aid of simple scripts. These scripts allow multiple Racadm commands to be automated for tasks such as the following:

- Managing usernames and passwords for a set of RACs
- Configuring RAC event management
- Updating RAC firmware across multiple RACs

Understanding Racadm installation and modes for configuring a RAC

The Racadm utility ships with Dell PowerEdge servers on the Dell OpenManage™ Systems Management CD and is also available at <http://support.dell.com>. Racadm can be installed and used on managed systems and management stations.

During the express installation of Dell OpenManage software, Racadm is installed by default (along with other RAC software components for managed systems) on Dell PowerEdge servers that contain RACs and are running supported versions of Microsoft® Windows® and Red Hat® Linux® operating systems. Racadm also can be installed on supported Windows-based management stations. For supported Novell® NetWare® operating systems, administrators must manually select to install Racadm and other RAC software components during managed-system installation.

Racadm supports three methods for configuring a RAC:

- Locally on the managed system, as an application running on the Microsoft MS-DOS® operating system, version 6.22 and higher; useful for pre-OS RAC configuration
- Locally on the managed system, as a Windows, Linux, or NetWare application
- Remotely from a management station, as a Windows application

Configuring a RAC in MS-DOS mode

Racadm commands can be executed in an MS-DOS environment. The MS-DOS mode enables administrators to perform scriptable configuration of the RAC prior to the deployment of an OS. To execute a Racadm command in MS-DOS mode, administrators must copy the racadm.exe file from the RAC directory on the Dell OpenManage Systems Management CD to a bootable MS-DOS disk.

Because Racadm runs on MS-DOS, administrators can use Racadm to configure RACs through third-party environments such as Microsoft Automated Deployment Services (ADS) and Altiris® server deployment and provisioning software.

Executing local and remote Racadm commands

Using the racadm command, administrators can enter subcommands to configure RAC properties. These RAC properties include user, session management, Simple Network Management Protocol (SNMP), and network and security settings. When administrators execute the Racadm subcommands, the Racadm utility sets or retrieves object property values from the RAC property database. These commands can be executed both locally on the managed system and remotely from a management station. To execute Racadm commands remotely, administrators must specify the IP address and a valid username and password for the RAC. The following example illustrates how an administrator can reset a RAC locally from a managed system and remotely from a management station.

From a managed system:

```
racadm racreset
```

From a management station:

```
racadm -r 192.168.100.001 -u root -p rootpassword  
racreset
```

Using a RAC configuration file to enable multiple RAC configurations

The Racadm command-line utility enables administrators to configure multiple RACs in a single step using a configuration file. A RAC configuration file is a simple text file, similar to an .ini file. Although the name of a RAC configuration file does not require a specific extension, throughout this article the .cfg extension is used.

The RAC configuration file contains a list of RAC objects and their associated property values (see Figure 1). Each object is categorized into a group that best describes it. For instance, the user management-related objects are arranged in the cfgUserAdmin group and the security settings-related objects are in the cfgRacSecurity group.

SAMPLE RAC CONFIGURATION FILE PARAMETERS

This representative set of RAC configuration file parameters can be configured and replicated across multiple RACs. Parameters are categorized by functional groups.

User settings

- Username (login)
- User password
- Administrator e-mail (enable/disable)
- Administrator e-mail address

Security configuration

- Secure Sockets Layer (SSL) handshake RSA key size
- Certificate Signing Request (CSR) common name and organization
- CSR locality, state, and country code
- CSR e-mail address
- Other authentication setting options

SNMP trap configuration

- IP address of trap destination
- Traps (enable/disable)
- SNMP trap community

RAC configuration files can be extremely useful in an IT environment where multiple RACs share similar settings. In such situations, administrators can create a configuration file containing property values for all objects that need to be replicated across multiple RACs in the environment. Administrators can then use a script to deploy those settings efficiently across the network to a set of RACs.

Creating a RAC configuration file

A RAC configuration file can be created three different ways, allowing administrators the flexibility to use the method best suited to their needs:

- Build the file from scratch by using the *Dell Remote Access Controller Racadm User's Guide* to obtain a list of all the available groups and objects
- Obtain a file from a RAC that is already installed and configured by using the `racadm getconfig` command
- Obtain a file using the `racadm getconfig` command and then customize the file as needed

The simplest way to create a configuration file is to execute the `racadm getconfig` command on a system whose RAC settings are to be replicated across multiple systems, and then edit the configuration file to remove system-specific settings like static IP addresses. If built from scratch, the configuration file must follow the parsing rules specified in the *Dell Remote Access Controller Racadm User's Guide*. This guide is available on the Product Documentation CD shipped with Dell PowerEdge servers and is also located at <http://docs.us.dell.com/docs/software/smdrac3/RAC/en/index.htm>.

Managing user authentication settings

The RAC supports a maximum of 16 administrators. Using the RAC configuration file, administrators can add new users or modify existing RAC user settings. These settings include usernames, user passwords, and user e-mail and page settings for receiving the RAC event alerts (see "Configuring SNMP event traps"). The objects for user settings are arranged in the `cfgUserAdmin` group. For example, to add or change a user password and enable e-mail alerts for a user named "John," an administrator would add or modify the following lines in the RAC configuration file under the `cfgUserAdmin` group:

```
cfgUserAdminUserName=John
cfgUserAdminPassword=1234
cfgUserAdminEmailEnable=1
cfgUserAdminEmailAddress=john@xyz.com
```

These commands will create a user named "John" if the user does not already exist, set the password to "1234," and enable

```
# Object Group "cfgUserAdmin"
#
[cfgUserAdmin]
# cfgUserAdminIndex=1
cfgUserAdminUserName=root
cfgUserAdminPrivilege=0
cfgUserAdminAlertFilterRacEventMask=0x300000
cfgUserAdminAlertFilterSysEventMask=0x77777
cfgUserAdminPageNumericEnable=0
cfgUserAdminPageNumericNumber=
cfgUserAdminPageNumericMessage=SE
cfgUserAdminPageNumericHangupDelay=0x0
cfgUserAdminPageAlphaEnable=0
cfgUserAdminPageAlphaNumber=
cfgUserAdminPageAlphaProtocol=8N1
cfgUserAdminPageAlphaBaudRate=0x4b0
cfgUserAdminPageAlphaCustomMsg=
cfgUserAdminPageAlphaModemConnectTimeout=0x3c
cfgUserAdminPageAlphaPagerId=
cfgUserAdminPageAlphaPassword=
cfgUserAdminEmailEnable=0
cfgUserAdminEmailAddress=
cfgUserAdminEmailCustomMsg=
cfgUserAdminPageModemInitString=AT+GCI=B5
cfgUserAdminPageModemPort=0x1
cfgUserAdminType=0x3

# Object Group "cfgTraps"
#
[cfgTraps]
# cfgTrapsIndex=1
cfgTrapsDestIpAddr=10.104.250.1
cfgTrapsEnable=1
cfgTrapsSnmpCommunity=
cfgTrapsFilterRacEventMask=0x300000
cfgTrapsFilterSysEventMask=0x77777

# Object Group "cfgRacSecurity"
#
[cfgRacSecurity]
cfgRacSecCsrKeySize=0x400
cfgRacSecCsrCommonName=
cfgRacSecCsrOrganizationName=
cfgRacSecCsrOrganizationUnit=
cfgRacSecCsrLocalityName=
cfgRacSecCsrStateName=
cfgRacSecCsrCountryCode=
cfgRacSecCsrEmailAddr=
cfgRacSecSslEnable=1
cfgRacSecVncInEncryptEnable=1
# cfgRacSecAuthLocalRacEnable=0x01
cfgRacSecAuthLocalOsEnable=1
```

Figure 1. A sample RAC configuration file showing objects and their property values

e-mail delivery of RAC event alerts to the e-mail address "john@xyz.com."

Configuring SNMP event traps

RACs can send SNMP traps to management consoles—for instance, a Dell OpenManage IT Assistant management station—when an event, such as the expiration of a server watchdog timer, occurs. Defined under the `cfgTraps` group, a maximum of 16 SNMP trap entries can be stored in the RAC management information base (MIB). Administrators can specify these SNMP trap settings by adding lines of code similar to the following in the RAC configuration file under the `cfgTraps` group:

```
cfgTrapsDestIpAddr=192.168.1.1
cfgTrapsEnable=1
cfgTrapsSnmpCommunity=public
```

These commands will enable SNMP traps for all RAC event alerts to be sent to the destination IP address of 192.168.1.1 using the community name "public." Each community name has a level of security (such as read-only or read-write); the server embeds the community name in an SNMP request for the receiving machine to check.

Using automation to replicate RAC configurations

A powerful feature of the `Racadm` utility is the capability to increase efficiency by automating the configuration of RACs. Once all the desired property values for the objects are specified in the RAC configuration file, administrators can use the `racadm config` command to deploy and replicate the settings specified in the RAC configuration file to multiple RACs in the environment.

The following sample command replicates the configuration contained in the `myconfig.cfg` RAC configuration file remotely to a RAC at the IP address 192.168.100.001:

```
racadm -r 192.168.100.001 -u root -p rootpassword
config -f myconfig.cfg
```

The following sample procedure demonstrates how to use a configuration captured from a model system to configure multiple RACs on the same network. In this example, the management station runs a Windows OS. Figure 2 shows a schematic view of this scenario.

1. Configure the source RAC on the model system.
2. Use the following command to capture the source RAC's settings in a file named `config_file.cfg`:

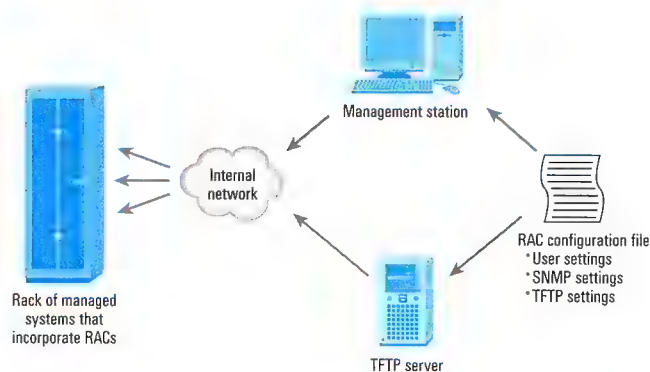


Figure 2. Configuration and deployment of multiple systems

```
racadm getconfig -f config_file.cfg
```

In these commands, "root" is the administrator username on the RACs and "rootpassword" is the administrator password. When a RAC is shipped, these values are set by default to "root" and "calvin," respectively.

3. If Dynamic Host Configuration Protocol (DHCP) is used to assign IP addresses, no IP address changes need to be made to the configuration file. However, if the source RAC is using a static IP address, remove or modify the static IP address information from `config_file.cfg` before configuring the other RACs.

The following command configures every RAC whose IP address is listed with the same configuration as the model system:

```
for %s in (198.62.20.2 198.62.20.3 198.62.20.4)
do racadm -r %s -u root -p rootpassword
config -f config_file.cfg
```

Refer to Microsoft Windows Help for information on the `for` command.

Alternatively, if a file listing the IP addresses for the RACs to be configured is available, insert the list dynamically by using the following command within a script:

```
for /f "usebackq" %s in ('type filename.lst')
do racadm -r %s -u root -p rootpassword
config -f config_file.cfg
```

Automating RAC firmware updates

Keeping up-to-date with the latest systems management software is critical for effective and efficient management of servers

in an enterprise environment. Performing RAC firmware updates for a significant number of PowerEdge servers can be a time-consuming task if performed one RAC at a time.

The Racadm utility simplifies this task by providing the `fwupdate` subcommand, which performs a firmware update in a single step by taking advantage of a RAC's ability to download firmware from a remote server running Trivial FTP (TFTP).

The following sample command can be used in a script to update RAC firmware on multiple RACs:

```
racadm fwupdate -g -u -a TFTP IP address
-f path/filename
```

In this example, the `-g` option instructs the RAC to download the firmware update file from a location specified by the `-f` option, using the TFTP server at the IP address specified by the `-a` option. The `-g` option loads the firmware update file into RAC memory, and the `-u` option instructs the RAC to perform the actual firmware update.


The following sample procedure demonstrates how to use a configuration captured from a model system to install the latest firmware on the specified RACs:

1. On a server running TFTP, which has an IP address of 192.68.0.1, download the latest firmware for the RACs from <http://support.dell.com>.
2. Place the RAC firmware file in the root directory of the TFTP server.
3. Update the firmware for each of the RACs using an approach similar to that used for distributing the RAC configuration:

```
For %s in (198.62.20.2 198.62.20.3 198.62.20.4)
do racadm -r
%s -u root -p rootpassword fwupdate -g -u -a
192.68.0.1 -f firmware_filename
```

Enabling efficient management of RACs using Racadm and scripting

By providing administrators with remote access to Dell PowerEdge servers, RACs can help increase server availability through early notification of potential or actual failures. In addition, RACs can improve administrator productivity by reducing travel time and the costs of managing remote servers. Also, RACs can increase server security by providing secure access that can be used for monitoring and controlling remote servers.

The Racadm command-line utility and native OS scripting simplifies the management of multiple RACs in an enterprise environment by enabling automation. Automation can help save time and reduce the need for IT resources by enabling administrators to deploy RAC configurations and updates efficiently across the network to a set of RACs. In this way, automation can streamline the deployment, configuration, and ongoing operation of RACs for IT departments that use a large number of Dell PowerEdge servers. 

Zain Kazim (zain_kazim@dell.com) is a test engineer in the Dell Enterprise System Test organization. His responsibilities include quality assurance of Dell enterprise products. Zain has a B.S. in Computer Science from Michigan State University.

Bala Beddhannan (bala_beddhannan@dell.com) is a test engineer in the Dell Enterprise System Test organization. Bala has a B.E. from Anna University, India, and an M.S. in Interdisciplinary Engineering from Texas A&M University.

Alan Daughetee (alan_daughetee@dell.com) is an engineering technician specialist in the Dell Enterprise System Test organization. He has more than four years of test experience with Dell enterprise products.

FOR MORE INFORMATION

To download Racadm:

<http://support.dell.com>

Dell Remote Access Controller Racadm User's Guide:

<http://docs.us.dell.com/docs/software/smdrac3/RAC/en/index.htm>

Share Your
Experience in
Dell Power Solutions

Dell Power Solutions is a peer-to-peer communication forum. We welcome subject-matter experts, end users, business partners, Dell™ engineers, and customers to share best-practices information. Our goal is to build a repository of solution white papers to improve the quality of IT.

Guidelines for submitting articles to *Dell Power Solutions* can be found at <http://www.dell.com/powersolutions>.

System Recovery

Using Windows Server 2003 on Dell PowerEdge Servers

By understanding different ways to recover a failed server, administrators can minimize interruptions to critical business processes. This article explores several powerful mechanisms for the recovery of Dell™ PowerEdge™ servers running the Microsoft® Windows Server™ 2003 operating system.

BY RANJITH PURUSH, NEFTALI REYES, AND EDWARD YARDUMIAN

Business interruptions and loss of productivity caused by system failure can be damaging and expensive. To help avoid unplanned outages, Microsoft has improved the robustness of the Windows Server™ 2003 operating system (OS), introducing enhanced system recovery features. At the same time, Windows Server 2003 continues to support the recovery options that were provided in the Microsoft® Windows® 2000 Server OS.

Dell also provides support for system recovery with both software- and hardware-based server management products. Dell™ remote access controllers (RACs) are one such category of hardware devices. RACs are supported in most Dell PowerEdge™ servers. Together, Microsoft and Dell recovery features can help provide smoother operation of Dell PowerEdge servers running Windows Server 2003.

Most recovery methods require advanced knowledge of the OS. Therefore, attempting to recover the system using basic recovery options such as Windows Safe-Mode Boot, Last Known Good Configuration, and Device Driver

Rollback—a feature new to Windows Server 2003—is recommended before proceeding to more advanced features. This article covers two advanced Microsoft Windows recovery options:

- **Emergency Management Services:** An advanced recovery option introduced in Windows Server 2003, Emergency Management Services (EMS)¹ provides powerful out-of-band management capabilities.
- **Automated System Recovery:** As a last resort when EMS functionality does not solve the problem, Automated System Recovery (ASR)² may be used to format the hard disks and recover the system from backed-up files.

Understanding in-band and out-of-band management

In-band management refers to the many mechanisms available to manage a system that has a fully functional OS and server hardware. For example, the Dell OpenManage™ systems management product suite offers in-band server

¹ Microsoft EMS should not be confused with the Dell OpenManage IT Assistant event management system (EMS), which monitors servers for specific events. For more information on event monitoring by IT Assistant, see "Understanding and Selecting Events to Monitor in IT Assistant" by Manoj Gujarathi in *Dell Power Solutions*, May 2002.

² Microsoft ASR should not be confused with automated server recovery (ASR), which is a hardware-based fault-tolerance feature. All ASR mentions in this article refer to Microsoft ASR.

management that enables administrators in an enterprise environment to manage hundreds of servers over the network from a single management console.

An in-band connection can be used only when the server is fully operational and accessible over the network. When these conditions cannot be met, out-of-band management methods must be used to manage the server remotely. Administrators can use out-of-band management when attempting to recover a system that has had a critical failure such as an unresponsive, or *hung*, OS.

An out-of-band connection uses a serial port, a modem, or a dedicated network interface card (NIC) together with specialized administration software—such as the EMS suite of applications—to communicate with the remote server. The out-of-band systems management approach does not replace in-band systems management; it simply enables administrators to quickly return a server to its fully functional state so they can once again control the server using in-band methods.

In a typical out-of-band management network, administrators control devices over a route separate from the main communication network. When in-band communication over a standard network path fails, administrators use the out-of-band network to connect to the servers, diagnose the problem, and effect repairs. Figure 1 shows a network that has both in-band and out-of-band connections, the latter implemented through serial connections.

An out-of-band connection through a serial console port relies on the most primitive OS services. As long as the kernel is functioning, the system can be accessed through the serial console port—even if the network stack or graphical user interface (GUI) is not operational. Because the console port delivers only text data, it offers good performance over low-bandwidth connections such as dial-up lines.

Using Windows EMS for out-of-band management

EMS is a powerful new out-of-band system recovery suite introduced in Windows Server 2003. EMS features, which involve multiple elements of the Windows Server 2003 OS, allow for system recovery through the server's serial console port even when the server is unavailable through the network because the operating system's network stack and user interface are not functional.

EMS support on Dell PowerEdge servers has two primary requirements:

- **BIOS support for console redirection:** To control and recover a server using out-of-band management, the server firmware must support console redirection; additional requirements may depend on an organization's specific

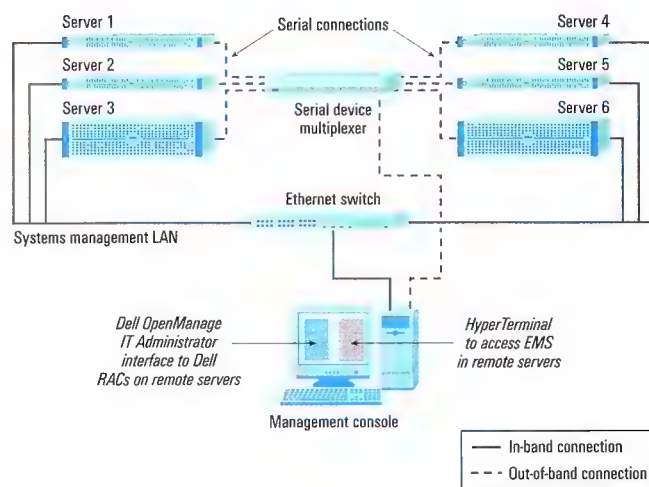


Figure 1. A server infrastructure that supports both in-band and out-of-band management

server management implementation. Fifth-generation PowerEdge servers (such as the PowerEdge 2500 and 2550) and newer PowerEdge servers (such as the PowerEdge 2650 and 6600) support console redirection by allowing the local display to be diverted to a remote console through serial ports. To enable or verify this feature, administrators can access the BIOS configuration page of a Dell PowerEdge server by pressing F2 immediately after the server starts booting, and then setting Console Redirection to "Enabled."

- **BIOS support for SPCR table:** With the exception of the PowerEdge 1550, all fifth-generation and newer PowerEdge servers support the Serial Port Console Redirection (SPCR) table. The SPCR table provides EMS with information about the out-of-band management port as well as related configuration details. If the Dell PowerEdge server does not support the SPCR table, administrators must provide the EMS configuration parameters and information about the out-of-band management port to the Windows Server 2003 OS by using the `bootcfg` command.³ The same command may be used to change the default configuration of EMS even if the server supports the SPCR table.

The main components of EMC are the two remote management consoles that are available only within EMS: the Special Administration Console (SAC) and the !Special Administration Console (!SAC). EMS also includes components that are standard features of the Windows Server 2003 OS, to which EMS has added console redirection capability.

³ For more information about using `bootcfg` to enable EMS, visit http://msdn.microsoft.com/library/default.asp?url=/library/en-us/ddtools/hh/ddtools/bootini_58xf.asp.

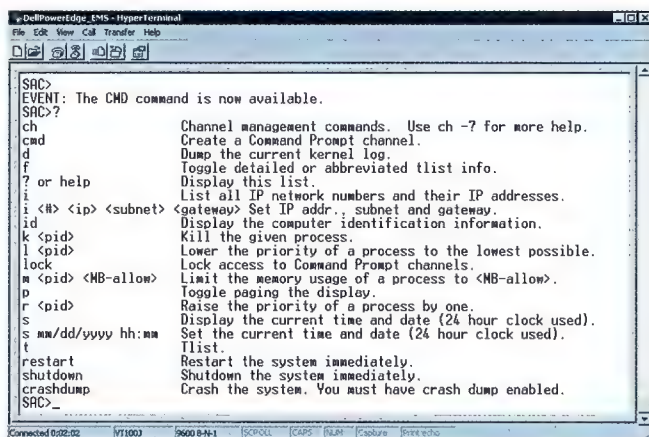


Figure 2. List of commands available in the EMS SAC

EMS supplies text-mode console redirection

EMS console redirection has been integrated into multiple aspects of the OS, including the setup loader, the text-based setup process, Remote Installation Services (RIS), the loader, and the Stop error handler. Support for console redirection in the setup loader, the text-based setup process, and RIS allows for remote installation of the OS using EMS. Meanwhile, console redirection support by the loader and the Stop error handler are crucial in aiding system recovery.

The console redirection feature allows administrators to view and manage a remote server. The Dell PowerEdge server BIOS supports text-mode console redirection during the server's power-on self-test (POST) until the OS begins to load. EMS supports text-mode console redirection as soon as Windows Server 2003 begins to load, and the feature remains available until the Windows GUI begins to load. This provides administrators with uninterrupted text-mode console redirection through an out-of-band management port beginning from server POST until the EMS SAC or !SAC session is available.

SAC enables OS management

SAC is the primary EMS command-line environment⁴ that allows administrators to access and control several Windows Server 2003 components. As a kernel-level function, SAC remains accessible even after high-level applications cease to respond. If a server is not responding because of a misbehaving process, administrators can use SAC to stop the errant process and even restart the server. By providing a limited set of very powerful commands, SAC can return the system to an in-band state. While in a SAC session, administrators can do the following:

- Set or view the IP address of the server
- Restart or shut down the server

- List all used and available resources (such as physical memory and kernel memory)
- List processes, kill processes, limit a process's memory usage, or change process priority
- Create command-prompt channels to access the file system, enabling administrators to run text-based applications such as the bootcfg command-line utility, replace system files, and copy files from CDs or floppy disks
- Invoke a crash dump that can be used for debugging
- Generate a Stop error that will create a memory dump file
- Dump the kernel log

For a complete list of SAC commands, use the `help` or `?` command from within the SAC session. This will display the command list as shown in Figure 2.

One of the most robust features of SAC is the provision for administrators to access the local file system. This capability is made possible by the Special Administration Console Helper service (sacsvr), which allows administrators to create up to eight simultaneous command-prompt channels. Access to each channel is secured and requires valid login credentials to a local or domain account. With access to the file system and even floppy disks, CDs, and network shares, the SAC command-prompt channels can be used to replace corrupted or missing system files. Although SAC will not remember the mapped network connections on the host server, SAC command-prompt channels can be used to access network shares as long as the TCP/IP stack on the host server is functional (see Figure 3).

If critical system files such as `pci.sys` or files that identify hardware necessary for the boot process are corrupted or missing,

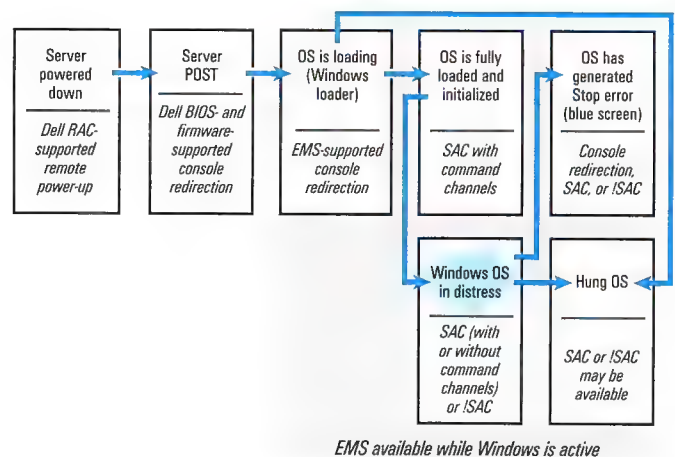


Figure 3. Availability of system recovery components on a Dell PowerEdge server when the Windows Server 2003 OS is in different states (arrows represent the transitions in OS state)

⁴ The SAC command-line environment is different from the standard command-line environment available in Windows operating systems and it offers different functionality.

some of the basic OS components including SAC may fail to load. When SAC fails to load, Windows Server 2003 attempts to make !SAC available.

!SAC provides auxiliary OS control

!SAC is an auxiliary EMS command-line environment hosted by Windows Server 2003 that also communicates through the out-of-band management port. The !SAC command-line environment is different from both the SAC and Windows Server 2003 command-line environments. When SAC fails to load or is not functioning, !SAC becomes available automatically.

The primary functions of !SAC are to redirect text from Stop error messages and allow administrators to restart the computer if SAC becomes unavailable. To this end, !SAC offers a subset of SAC commands, which can do the following:

- Restart the server
- Display computer identification information and all log entries
- Display Stop error message explanatory text

If missing or corrupted critical system files prevent even !SAC from loading, administrators should use alternative methods to replace key system files.⁵

Terminal emulation software establishes SAC and !SAC sessions on remote servers

The serial port is the most common out-of-band management port used by EMS, and PowerEdge servers continue to provide a legacy serial port that is supported by Windows Server 2003. Industry-standard conventions for terminal emulation such as VT-UTF8 and VT-100 enable remote EMS consoles to send commands to the server through a serial connection. With its support for enhanced escape sequences and language localization, VT-UTF8 is the preferred terminal type for viewing EMS output. However, if VT-UTF8 emulation is not available or supported by the terminal emulation software, VT-100+ or VT-100 emulation can be used.⁶ Dell PowerEdge servers currently support only VT-100 emulation for console redirection.

A terminal emulator such as the Hilgraeve HyperTerminal® utility, which supports VT-UTF8 and other terminal control standards, can be used for viewing EMS output sent through the server's serial port. The Hilgraeve HyperTerminal utility ships with most Windows client and server operating systems, including Windows

Microsoft EMS	Dell RAC	Microsoft ASR
<ul style="list-style-type: none"> • Support out-of-band management only • Allow remote power-down • Allow remote reboot • List and kill processes • List used and available resources • Access file system, local external storage media, and network shares • Configure IP address • Redirect Stop error messages • Dump kernel log and invoke crash dump 	<ul style="list-style-type: none"> • Support out-of-band and in-band management • Allow remote power-up • Allow remote power-down • Support graphic and text-mode console redirection if OS is functional • Access hardware log, POST log, and boot log • Monitor system health and alert administrators to potential hardware problems • Perform remote floppy boot function 	<ul style="list-style-type: none"> • Offer alternative recovery option when all other recovery operations fail • Require backup and restore of system volumes • Require administrators to proactively make periodic backups using the ASR utility • Support backup and restore of system state only (not a full data backup utility)

This listing includes only features that may be useful for system recovery, and is not representative of the full feature set for each component.

Figure 4. Components useful for system recovery of Dell PowerEdge servers running Windows Server 2003

Server 2003. The connection from the host client system that runs the terminal emulation software to the target server should be through a null modem cable. For information on configuring a HyperTerminal session to connect to the EMS console and on accessing EMS SAC and !SAC using HyperTerminal, see "Using HyperTerminal to send EMS commands to the server."

Exploring alternative system recovery options

The EMS implementation in Windows Server 2003 allows administrators to attempt out-of-band system recovery when the OS or the software stack on the server might have caused a system failure. With the Dell OpenManage systems management suite, Dell extends the capabilities of administrators to troubleshoot hardware issues by supporting in-band and out-of-band management. Dell OpenManage Server Administrator—which can monitor a server's health including voltage, temperature, and cooling fan status—can be configured to alert administrators of any potential problems.⁷

Figure 4 summarizes the features of each system recovery option discussed in this article and how those features may contribute to the recovery of PowerEdge servers running Windows Server 2003.

Windows EMS enhances the functionality of a Dell RAC

On many PowerEdge servers Dell provides a RAC, which is a hardware-based remote console that is completely independent of

⁵ The Windows loader (NTLDR) makes available a concise version of !SAC that allows administrators to restart the server. This version of !SAC is automatically made available to either the remote or local console by the loader when it cannot load the kernel.

⁶ For more information about selecting client terminal software for EMS and supported conventions and escape sequences, visit http://www.microsoft.com/technet/treeview/default.asp?url=/technet/prodtechnol/windowsserver2003/proddocs/standard/EMS_VT100_conventions.asp.

⁷ For more information on the Dell OpenManage suite, visit <http://support.dell.com/systemdocumentation/index.aspx?category=6,111>.

USING HYPERTERMINAL TO SEND EMS COMMANDS TO THE SERVER

The following two procedures indicate how to connect to Microsoft EMS and access EMS SAC and !SAC using the Hilgraeve HyperTerminal utility.

Configuring HyperTerminal to connect to EMS using a serial port

Prerequisites:

- Enable console redirection on the target server system BIOS by pressing F2 immediately after the server starts booting, and then setting Console Redirection to "Enabled."
- Connect a null modem cable between the client system's serial COM port and the target server's serial COM port.

1. On a client computer, open a new HyperTerminal session from the Start menu by selecting Start > Programs > Accessories > HyperTerminal. If HyperTerminal is not available in the Accessories list, locate it by selecting Start > Programs > Accessories > Communications > HyperTerminal.

Note: Although the HyperTerminal utility is available in Microsoft Windows Server 2003, it is not installed by default. It can be installed manually by enabling it in the Add/Remove Windows Component applet. To do so, go to Start > Settings > Control Panel > Add/Remove Programs > Add/Remove Windows Components > Accessories and Utilities > Communication > HyperTerminal.

2. In the Connection Description dialog box, enter a name for the connection.
3. In the Connect Using field on the Connect To dialog box, select the COM port on the client that has been connected to the server.
4. Choose the following in the COM Port Properties dialog box:
 - Bits per second: 9,600
 - Data bits: 8

- Parity: None
- Stop bits: 1
- Flow control: Hardware

The HyperTerminal serial communications program is now ready to use.

Accessing EMS SAC and !SAC using HyperTerminal

1. Complete the previous procedure to configure a HyperTerminal session between the client and target systems.

2. Power on the target server.

The Dell server firmware will redirect the power-on self-test (POST) information for the server to the HyperTerminal session. After the POST, when the OS begins to load, Dell firmware console redirection will end and EMS console redirection will begin. The following message will appear on the HyperTerminal session window: "Computer is booting, SAC started and initialized."

At this time, most SAC commands are accessible, but the command-prompt channels are not yet available because the SAC services have not yet completely registered. SAC is fully operational when it displays the following message: "EVENT: The CMD command is now available."

3. If the Redirection After Boot option has been selected in the server system BIOS, the SAC messages will be blinking. Press Esc + Tab to stop the blinking.
4. Use the `help` or `?` command at the SAC prompt to display a list of supported SAC commands.
5. Use the Esc and Tab keys to toggle between the command channels and the SAC console.

the server firmware and OS.⁸ The RAC's dedicated network port offers an Ethernet-based out-of-band management option. Typically, when a server hangs, administrators must physically press the power button to power cycle the server. The capability of the RAC to remotely cold boot, restart, and shut down servers that do not have a functional OS can greatly enhance the power cycling capabilities of Windows recovery systems. When used in conjunction with the EMS available in Windows Server 2003, a RAC can help provide a powerful and robust recovery system.

All the latest Dell servers ship with the Dell Embedded Remote Access (ERA) controller or support the Embedded Remote Access

Option (ERA/O). Older systems support variations of the ERA controller such as the Dell Remote Access Card III (DRAC III) and DRAC III/XT, which provide a subset of the functionality that the ERA controller offers.⁹

ASR recovers a failed system and restores system state

ASR is an advanced option of the Windows Server 2003 Backup Tool that allows administrators to recover a failed system and restore it to the configuration that was most recently backed up. In Windows Server 2003, ASR takes the place of the Emergency Repair Disk feature found in earlier Microsoft server operating

⁸ In this article, Dell Remote Access Card III (DRAC III), DRAC III/XT, Embedded Remote Access (ERA), and Embedded Remote Access Option (ERA/O) are collectively referred to as RACs. When information applies only to a specific RAC, it is identified explicitly.

⁹ For more information on Dell RACs, visit <http://support.dell.com/docs/software/smdrac3/RAC/en/is/racugc1.htm#28099>. To access the user guides for the different RACs, visit <http://support.dell.com/docs/software/smdrac3/index.htm>.

systems. *Note:* It is highly recommended that ASR be used only after all other system recovery methods have been attempted and the only remaining option is to reformat the disk and reinstall the OS.

The advanced recovery options discussed thus far are tools that can be used to attempt to recover a failed system without requiring any prerequisites such as a system backup. With the ASR Backup utility, administrators can initiate periodic backups of server system state and then, with the ASR Restore utility, use these backup files to attempt to restore system state when a server encounters a critical failure.

Without a backup, ASR Restore will not be able to restore a failed system. In such cases, the only option to recover the system is for administrators to reinstall the OS and reconfigure all physical storage manually.

ASR allows for the restoration of system state, critical files on the system, and boot partitions. System state includes the following:

- Boot files and system files
- Files protected by Windows File Protection (WFP)
- The registry
- Performance counter configuration information
- The Component Object Model+ (COM+) class registration database
- The certificate services database (if the server is a Microsoft Certificate Server)
- Microsoft Active Directory® directory service database
- The sysvol directory (if the server is a domain controller)
- Cluster database information (if the server is a node in a cluster)
- Internet Information Services (IIS) metabase (if the server has IIS installed)


Backing up user data is critical and should be implemented as a separate process from ASR, which is not a data backup utility. Although ASR can restore data files that are in the system or boot volumes, it does not restore data files that reside in other logical or physical partitions or volumes. However, if the disks that host the data volume are not corrupted, they may be accessible once ASR restoration is complete.

An ASR restoration typically follows these steps:

- Rebuild the critical volumes (volumes that host the OS) using the information stored in a floppy disk that was created during ASR backup.
- Perform a simple installation of Windows Server 2003 using the Windows Server 2003 installation CD.
- Begin the restoration automatically from the media on which the ASR backup was created.

For step-by-step instructions on how to use ASR for backup and restore operations, visit *Dell Power Solutions* online at http://www.dell.com/magazines_extras.

Developing more robust recovery mechanisms

With the introduction of EMS and enhanced ASR features, Windows Server 2003 offers IT organizations robust system recovery options. Dell PowerEdge servers provide the required hardware and software stack that further expedites system recovery. In the future, the Dell OpenManage systems management suite will introduce enhanced support for EMS and other recovery options that will allow administrators to use more robust and standardized Ethernet-based out-of-band mechanisms to handle system recovery. 

Ranjith Purush (ranjith_purush@dell.com) is a systems engineer in the Server Operating Systems Engineering Department at Dell. His current areas of focus include virtualization software and performance benchmarking. Ranjith has an M.S. in Electrical and Computer Engineering from The University of Texas at Austin.

Neftali Reyes (neftali_reyes@dell.com) is a systems engineer in the Server Operating Systems Engineering Department at Dell. He has a B.S. in Computer and Information Systems Management from Park University in Missouri, and an associate's degree in Electronics Technology from Austin Community College. He is a Microsoft Certified Systems Engineer (MCSE).

Edward Yardumian (edward_yardumian@dell.com) manages the Operating Systems Engineering and Certification teams in the Dell Enterprise Product Group. Previously, Edward led engineering projects for clustering and next-generation PowerEdge servers. He has published numerous articles and has patents pending on cluster computing and scalable solutions.

FOR MORE INFORMATION

Disaster recovery:

http://www.microsoft.com/technet/treeview/default.asp?url=/technet/prodtechnol/windowsserver2003/proddocs/entserver/concepts_recovery.asp

Microsoft Emergency Management Services:

http://www.microsoft.com/technet/treeview/default.asp?url=/technet/prodtechnol/windowsserver2003/proddocs/entserver/ems_topnode.asp

<http://www.microsoft.com/whdc/hwdev/platform/server/headless/default.mspx>

Microsoft Automated System Recovery:

http://www.microsoft.com/technet/treeview/default.asp?url=/technet/prodtechnol/windowsserver2003/proddocs/entserver/asr_overview.asp

Dell remote access controllers:

<http://support.dell.com/docs/software/smdrac3/RAC/en/is/racugc1.htm#28099>

<http://support.dell.com/docs/software/smdrac3/index.htm>

http://www1.us.dell.com/content/topics/global.aspx/power/en/ps2q02_bell?c=us&cs=555&l=en&s=biz

Simplifying Linux Management with Dynamic Kernel Module Support

At times, administrators may need newer drivers than the ones found in the Linux® operating system kernel. Dynamic Kernel Module Support, a software project created by the Dell™ Linux Engineering team, efficiently decouples driver releases from kernel releases, helping to provide an orderly method for distributing the latest drivers even when they are not yet merged into the Linux kernel.

BY GARY LERHAUPT AND MATT DOMSCH

As the Linux® operating system (OS) gains a deeper foothold in enterprise environments, system administrators have become increasingly concerned about the management of Linux kernel modules. In the best-case scenario, every driver needed to run every piece of system hardware would come precompiled with the Linux kernel. In practice, however, hardware drivers often are released separately from the kernel, and updates for drivers native to the kernel also are released independently, superseding the drivers within the latest kernel version.

Until the ultimate goal of pushing all driver modifications back into the kernel is met, Dynamic Kernel Module Support (DKMS), a software project created by the Dell™ Linux Engineering team, can help administrators add, build, install, remove, and track Linux kernel modules. DKMS aims to create a standardized framework for collecting driver source code, building this source code into loadable compiled-module binary files, and then installing and uninstalling these modules into the Linux kernel as needed. In addition, DKMS provides powerful features for managing and maintaining modules across multiple systems and keeping track of which module version is installed on which kernel version.

By creating a separate framework for driver source code and the module binary files that are compiled from that source code, DKMS efficiently decouples driver releases from kernel releases. Decoupling driver and kernel releases permits administrators to update drivers on existing kernels in an orderly and supportable manner as soon as they are available. Thus, DKMS serves as a stopgap, providing a way to distribute the latest driver updates until the source code can be merged back into the kernel.

In addition, DKMS streamlines the process of compiling from source code. Rebuilding RPM™ (Red Hat®

Package Manager) source packages can be time-consuming and problematic. DKMS helps simplify Linux development by creating a single executable that can be called to build, install, or uninstall modules.

Further, DKMS makes configuring modules on new kernels particularly

In the best-case scenario,
every driver needed
to run every piece of
system hardware would
come precompiled
with the Linux kernel.

easy for less-experienced Linux developers: The modules to be installed can be based solely on the configuration of a kernel that was previously running. In production environments, this represents an immediate advantage. For example, using DKMS, IT managers no longer have to choose between a predefined solution stack or the security enhancements of a newer kernel.

DKMS has two target audiences: developers who maintain and package drivers, and system administrators. This article focuses on DKMS from the system administrator perspective of using DKMS to simplify Linux enterprise computing management.¹

Understanding basic DKMS commands

Before exploring the uses of DKMS, it is helpful to understand the life cycle by which DKMS maintains kernel modules. Figure 1 represents each potential state for a module—Added, Built, and Installed—and each arrow indicates a DKMS action that can be used to switch between the various states. The sections that follow examine each of these DKMS actions further.

Most importantly, DKMS was designed to work with RPM. Using DKMS to install a kernel module often can be as easy as installing a DKMS-enabled module RPM, because module packagers can use DKMS to add, build, and install modules within RPM packages. Wrapping DKMS commands inside an RPM package preserves the benefits of RPM—package versioning, security, dependency resolution, and package distribution methodologies—while DKMS handles the work that RPM does not: the versioning and building of individual kernel modules. Of course, DKMS works just as well when not used in conjunction with RPM, so it is important to understand how to use these basic commands to fully leverage the capabilities of DKMS.

Add command adds a module and module version to the tree

DKMS manages kernel modules at the source-code level. First, the module source code must be located in the directory `/usr/src/module-module-version` on the build system. A `dkms.conf` file with appropriately formatted directives also must reside within this configuration file to tell DKMS where to install the module and how to build it. The `dkms.conf` file should come

Decoupling driver
and kernel releases
permits administrators
to update drivers on
existing kernels in an
orderly and supportable
manner as soon as
they are available.

from the module packager and be included with the module source code. Once these two requirements have been met and DKMS has been installed on a system, administrators can begin using DKMS by adding a module and module version to the DKMS tree. For example:

```
dkms add -m megaraid2 -v 2.00.9
```

This sample `add` command would add `megaraid2/2.00.9` to the already existing `/var/dkms` tree, leaving the module in an Added state.

Build command compiles the module

Once in the Added state, the module is ready to be built using the DKMS `build` command. The `build` command requires that the proper kernel source code be located on the system in the `/lib/module/kernel-version/build` directory. The `make` command that is used to compile the module is specified in the `dkms.conf` configuration file. The following sample `build` command continues the `megaraid2/2.00.9` example:

```
dkms build -m megaraid2 -v 2.00.9 -k 2.4.21-4.ELsmp
```

The `build` command compiles the module but stops short of installing it. As this example indicates, the `build` command expects a kernel-version parameter. If this kernel name is left out, it assumes the currently running kernel. The `build` command also can build modules for kernels that are not currently running; this functionality is provided through use of a kernel preparation subroutine that runs before any module build is performed. The subroutine ensures that the module being built is linked against the proper kernel symbols.

In this example, successful completion of a build creates the `/var/dkms/megaraid2/2.00.9/2.4.21-4.ELsmp` directory as well as the log and module subdirectories within this directory. The log directory holds a log file of the module make and the module directory holds copies of the resultant `.o` binary files that were compiled.

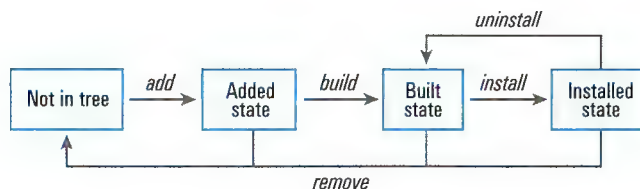


Figure 1. The DKMS life cycle

¹ For a detailed discussion about creating and developing DKMS-enabled module packages from the developer's perspective, see "Exploring Dynamic Kernel Module Support" in *Linux Journal*, September 2003, <http://www.linuxjournal.com/article.php?id=6896>.

Install command copies the compiled module binary files to the kernel tree

Upon completion of a build, the module can be installed on the kernel for which it was built. The `install` command copies the compiled module binary files to the correct location in the `/lib/modules` tree as specified in the `dkms.conf` file. If a module by that name already resides in that location, DKMS saves the existing module in the `/var/dkms/module-name/original_module` directory. This process helps ensure that the older module can be put back into place if, at a later date, the newer module is uninstalled. A sample `install` command is as follows:

```
dkms install -m megaraid2 -v 2.00.9 -k 2.4.21-4.ELsmp
```

In this example, if an original `megaraid2` module existed within the `2.4.21-4.ELsmp` kernel, it would be saved to `/var/dkms/megaraid2/original_module/2.4.21-4.ELsmp`.

Uninstall and remove commands expunge modules to differing degrees

The DKMS life cycle also enables administrators to uninstall or remove a module from the tree. The `uninstall` command removes the installed module and, if applicable, replaces it with the original module. When multiple versions of a module are located within the DKMS tree, if one version is uninstalled, DKMS does not try to determine which of these other versions to put in its place. Instead, if a true “original_module” was saved from the very first DKMS installation, it will be put back into the kernel and all other versions of that module will be left in the Built state. A sample `uninstall` command is as follows:

```
dkms uninstall -m megaraid2 -v 2.00.9
-k 2.4.21-4.ELsmp
```

Again, if the kernel-version parameter is unset, the currently running kernel is assumed. However, this same behavior does not occur with the `remove` command. Although the `remove` and `uninstall` commands are similar, some important differences exist. The `remove` command uninstalls, but also is used to clean the DKMS tree. If the module version being removed is the last instance of that module version for all kernels on a system, after the `uninstall` portion of the `remove` command completes, the `remove` command will physically delete all traces of that module from the DKMS tree. That is, when the `uninstall` command completes, modules are left in the Built state; when the `remove` command completes, an administrator would have to start over from the `add` command before being able to again use the module with DKMS. Two sample `remove` commands are shown here:

```
dkms remove -m megaraid2 -v 2.00.9
-k 2.4.21-4.ELsmp
dkms remove -m megaraid2 -v 2.00.9 --all
```

DKMS serves as a stopgap, providing a way to distribute the latest driver updates until the source code can be merged back into the kernel.

The first sample command would uninstall the module; if this module and module version were not installed on any other kernel, the command would remove the module from the DKMS tree altogether. If, however, `megaraid2/2.00.9` module and module version also were installed on the `2.4.21-4.ELhugemem` kernel, the first `remove` command would leave the module alone, and thus it would remain intact in the DKMS tree. Because the second sample command contains the `--all`

parameter, not the `-k kernel` parameter, the second command would uninstall all versions of the `megaraid2/2.00.9` module from all kernels and then completely expunge any references of `megaraid2/2.00.9` from the DKMS tree.

Extending DKMS functionality with auxiliary commands and services

The `add`, `build`, `install`, `uninstall`, and `remove` commands—which correlate to the DKMS life cycle—are the fundamental DKMS commands. The auxiliary DKMS functionality discussed in this section extends and improves upon the capabilities of these basic commands.

Status command returns data about modules currently located in the tree

DKMS also includes a fully functional `status` command that returns information about the modules and module versions currently located in the tree. The specificity of the information returned depends on which parameters are passed to the `status` command. If no parameters are set, this command will return all information found. Each status entry will return output—added, built, or installed—to indicate the state; and if an original module has been saved, this information also will be displayed. Several sample `status` commands are shown here:

```
dkms status
dkms status -m megaraid2
dkms status -m megaraid2 -v 2.00.9
dkms status -k 2.4.21-4.ELsmp
dkms status -m megaraid2 -v 2.00.9
-k 2.4.21-4.ELsmp
```

Match command applies module configurations from one kernel to another

Another major feature of DKMS is the `match` command. The `match` command takes the configuration of a DKMS-installed module for one kernel and applies the same configuration to another kernel.

When the `match` command completes, the same module and module versions that were installed for one kernel are installed on the other kernel. This is helpful to administrators who are upgrading from an existing kernel to a newer kernel, but would like to keep the same DKMS modules in place for the new kernel. A sample `match` command is as follows:

```
dkms match --templatekernel 2.4.21-4.ELsmp
-k 2.4.21-5.ELsmp
```

As shown in the preceding example, the `--templatekernel` parameter is the kernel on which the configuration is based, while `-k` is the kernel upon which the configuration is instated.

Dkms_autoinstaller service automatically installs a designated module

The `dkms_autoinstaller` service is similar in behavior to the `match` command. This service is installed in the `/etc/init.d` directory as part of the DKMS RPM. If an `autoinstall` parameter is set within the `dkms.conf` configuration file in a module, that module is eligible for the `dkms_autoinstaller` service to automatically, upon booting, build it into a new kernel. When the administrator later boots a system into a new kernel, the `dkms_autoinstaller` service will then automatically build and install modules designated for use with this service.

Mkdriverdisk command creates a driver disk image

The final auxiliary DKMS command is `mkdriverdisk`. As its name suggests, the `mkdriverdisk` command builds modules to create a driver disk image for use in distributing updated drivers to Linux installations. A sample `mkdriverdisk` command might look like this:

```
dkms mkdriverdisk -d redhat -m megaraid2
-v 2.00.9 -k 2.4.21-4.ELBOOT
```

Currently, the only supported distribution driver disk format is Red Hat. For more information on the extra necessary files and their required formats for DKMS to create Red Hat driver disks or for general information on Red Hat driver disks, see <http://people.redhat.com/dledford>. When creating driver disks with DKMS, administrators should place these files in a subdirectory underneath the module source directory: for example, `/usr/src/module-module-version/redhat_driver_disk`.

Managing multiple systems using `mktarball` and `ldtarball` commands

As the preceding examples demonstrate, DKMS provides a simple mechanism to build, install, and track driver updates. This functionality not only is applicable to stand-alone machines, but also is useful for IT departments that administer multiple similar servers. The DKMS `mktarball` and `ldtarball` commands enable organizations

DKMS helps simplify
Linux development
by creating a single
executable that can be
called to build, install,
or uninstall modules.

having a compiler and kernel source on only one system—a master build system—to deploy a new driver to multiple additional systems.

The `mktarball` command packages copies of each `.o` binary file from the module directory that was compiled using the DKMS `build` command into a compressed tar file. This compressed tar file may then be copied to each target system. Administrators can use the DKMS `ldtarball` command to load the compressed tar files into a DKMS tree, leaving each module in the Built state, ready to be installed. The `mktarball` and `ldtarball` commands keep administrators from having to install both kernel source code and compilers on every target system.

The following example assumes that an administrator has built the `megaraid2` driver, version 2.00.9, for two different kernel families—2.4.20-9 and 2.4.21-4.EL—on a master build system:

```
# dkms status
megaraid2, 2.00.9, 2.4.20-9: built
megaraid2, 2.00.9, 2.4.20-9bigmem: built
megaraid2, 2.00.9, 2.4.20-9BOOT: built
megaraid2, 2.00.9, 2.4.20-9smp: built
megaraid2, 2.00.9, 2.4.21-4.EL: built
megaraid2, 2.00.9, 2.4.21-4.ELBOOT: built
megaraid2, 2.00.9, 2.4.21-4.ELhugemem: built
megaraid2, 2.00.9, 2.4.21-4.ELsmp: built
```

To deploy this version of the driver to several systems without rebuilding from the source code each time, an administrator can use the `mktarball` command to generate two compressed tar files—one for each kernel family:

```
# dkms mktarball -m megaraid2 -v 2.00.9
-k 2.4.21-4.EL,2.4.21-4.ELsmp,2.4.21-4.ELBOOT,
2.4.21-4.ELhugemem
```

```
Marking /usr/src/megaraid2-2.00.9 for archiving...
Marking kernel 2.4.21-4.EL for archiving...
Marking kernel 2.4.21-4.ELBOOT for archiving...
Marking kernel 2.4.21-4.ELhugemem for archiving...
Marking kernel 2.4.21-4.ELsmp for archiving...
Tarball location: /var/dkms/megaraid2/2.00.9/
tarball/megaraid2-2.00.9-kernel2.4.21-4.EL-
kernel2.4.21-4.ELBOOT-kernel2.4.21-4.ELhugemem-
kernel2.4.21-4.ELsmp.dkms.tar.gz
Done.
```


When one large compressed tar file that contains modules for both families is preferred, administrators can omit the `-k` parameter and kernel list; DKMS then will include a module for every kernel version found.

After creating one or more compressed tar files, administrators should run the `status` command to ensure that the target DKMS tree does not already contain the modules to be loaded:

```
# dkms status
Nothing found within the DKMS tree for this
status command.
If your modules were not installed with DKMS,
they will not show up here.
```

Next, the compressed tar file can be renamed, if desired, and copied to each of the target systems using any mechanism. The compressed tar file is then loaded on the target system:

```
# dkms ldrtarball --archive=megaraid2-2.00.9-
kernel2.4.21-4.EL-kernel2.4.21-4.ELB00T-
kernel2.4.21-4.ELhugemem-kernel2.4.21-4.ELsmp.dk
ms.tar.gz
Loading tarball for module: megaraid2 / version:
2.00.9
Loading /usr/src/megaraid2-2.00.9...
Loading /var/dkms/megaraid2/2.00.9/2.4.21-4.EL...
Loading /var/dkms/megaraid2/2.00.9/2.4.21-
4.ELB00T...
Loading /var/dkms/megaraid2/2.00.9/2.4.21
-4.ELhugemem...
Loading /var/dkms/megaraid2/2.00.9/2.4.21-4.ELsmp...
Creating /var/dkms/megaraid2/2.00.9/source symlink...
```


The DKMS `ldrtarball` command leaves modules in the Built state, not the Installed state. Administrators should verify both that the modules are present and that they are in the Built state:

```
# dkms status
megaraid2, 2.00.9, 2.4.21-4.EL: built
megaraid2, 2.00.9, 2.4.21-4.ELB00T: built
megaraid2, 2.00.9, 2.4.21-4.ELhugemem: built
megaraid2, 2.00.9, 2.4.21-4.ELsmp: built
```

The preceding steps must be repeated for each kernel version into which modules are to be installed.

Simplifying administration and increasing system stability with DKMS

DKMS can simplify Linux system administration by providing a versioning framework for installing driver modules on kernels. DKMS integrates with RPM for package distribution and installation, facilitates OS installation on new hardware, and helps maintain driver consistency across multiple servers. By enabling deployment of driver updates independent of kernel updates, DKMS reduces the scope of change for configuration management, and thus can help increase system stability.

DKMS is licensed under the GNU General Public License (GPL). Interested parties may contribute to its development by signing up for the `dkms-devel@lists.us.dell.com` mailing list located at <http://lists.us.dell.com>. DKMS can be downloaded from <http://linux.dell.com/dkms>. 

Gary Lerhaupt (gary_lerhaupt@dell.com) is a software engineer on the Linux Engineering Team of the Dell Product Group, and is the author of the Dynamic Kernel Module Support project. Gary is a Red Hat Certified Engineer (RHCE) and has a B.S. in Computer Science and Engineering from The Ohio State University.

Matt Domsch (matt_domsch@dell.com) is a lead and senior engineer on the Linux Engineering Team of the Dell Product Group, which tests Linux on all Dell PowerEdge™ servers. Matt has an M.S. in Computer Science from Vanderbilt University and a B.S. in Computer Science and Engineering from the Massachusetts Institute of Technology. His primary areas of interest include networking and operating systems.

FOR MORE INFORMATION

DKMS project home page:
<http://linux.dell.com/dkms>

DKMS mailing list:
<http://lists.us.dell.com>

DKMS information for driver maintainers:
<http://www.inuxjournal.com/art.cle.php?sid=6896>

Red Hat driver disk reference:
<http://people.redhat.com/dledford>



Configuring and Managing Software RAID with Red Hat Enterprise Linux 3

Thanks to dramatic advances in processing power, software RAID implementations are now a viable alternative to hardware-based RAID. By providing a mature software RAID layer and several management tools, the Red Hat® Enterprise Linux® 3 operating system can help system administrators build effective, cost-efficient software RAID implementations.

BY JOHN HULL AND STEVE BOLEY

Hardware-based RAID controllers have long been the preferred method for implementing RAID storage because they offload the management of RAID arrays onto a separate processor, freeing precious system CPU cycles for other tasks. However, the price/performance ratio of CPUs has decreased, making software-based RAID a viable alternative for system administrators in Linux®-based environments. The 2.4 kernel of the Linux operating system (OS) and its mature software RAID layer and management tools, particularly in Red Hat® Enterprise Linux 3, enable administrators to build an inexpensive RAID implementation.

Understanding software RAID in Linux environments

The Linux 2.4 kernel provides software-based RAID through the md device-driver layer, which sits on top of the storage controller device drivers. Because it is device-independent, the md device-driver layer, or *Linux RAID layer*, can work with all types of storage devices including SCSI and IDE. RAID-0 (striping), RAID-1 (mirroring), and RAID-5 (striping with parity) are supported in this kernel-level RAID implementation. Device nodes for md are denoted as /dev/mdx, where *x* is a number from 0 to 15.

For software RAID, the `Linux raid autodetect` partition type is fd, which is an ID in the same manner that 83 is the type for ext3 partitions, 8e is the type for Logical Volume Manager (LVM) partitions, and 82 is the type for Linux swap partitions. When the Linux kernel boots, it automatically detects fd partitions as RAID partitions and starts the RAID devices. All kinds of partition types can be used with the Linux OS, including FAT16, FAT32, and even IBM® AIX® partitions, but fd is preferred because it does not require system administrators to start the RAID devices manually during boot as other partition types do.

The most useful RAID levels for enterprise storage usually are RAID-1 and RAID-5. Although this article focuses on creating and managing RAID-1 arrays in a Linux environment, much of the information also is applicable to RAID-5.

Creating RAID arrays during Linux OS installation

The easiest and most reliable method for configuring software RAID occurs during a new OS installation. For Red Hat Enterprise Linux 3, the Disk Druid tool provides a simple interface to define and create software RAID

configurations. When creating software RAID partitions and RAID-1 md devices, administrators must ensure that the system is configured correctly. Common best practices include:

- Create partitions that will correspond to the same *sdxy* device in the same order on each hard drive, where *x* changes with each drive but *y* remains the same for ease of administration in case of a drive failure. For example, *sda1* and *sdb1* (the first partitions on hard drives *sda* and *sdb*) both have a size of 100 MB. When administrators tag them with the *fd* partition type during installation, after creating a RAID device, these partitions become parts of device *md0*.
- Define precisely where each partition will reside, instead of letting Disk Druid determine the disk location. Disk Druid sometimes scatters partitions across disks, which can create problems for administrators if a disk must be replaced and the partitions rebuilt. To help determine where partitions reside, administrators can designate certain partitions as primary, which can help keep *sda1* through *sda3* and *sdb1* through *sdb3* aligned as *md0*, *md1*, and *md2*. This practice can help ease administration and recoverability.
- After creating matching partitions, create the *md* device for those partitions before defining the next set of partitions and *md* devices. This practice enables administrators to keep track more easily of which partitions match each other.
- Mirror all the partitions on the hard drives—including */boot*, */swap*, and so forth—to perform true RAID-1 mirroring. This practice helps ensure that all data on the system is backed up and can be restored if one drive fails.

Administrators then should complete the following steps in Disk Druid to create each software RAID device:

1. To create a software RAID partition, click the RAID button and then select “Create a software RAID partition.” For the file system type, select “software RAID.”
2. Ensure that only one drive is selected for the partition (for example, “*sda*” or “*hda*”), and make the desired configuration selections.
3. Repeat steps 1 and 2 but select the second hard drive (for example, “*sdb*” or “*hdb*”) on which to create a RAID partition.
4. Click the RAID button again and select “Create a RAID device” when prompted. Choose the mount point, file system type, RAID device, and RAID level for this set of RAID partitions.

Prepping the system for drive failure

After configuring the RAID devices and installing the OS, administrators should prepare the system so that the RAID configuration can easily be restored to a failed drive. This process involves making a

The price/performance ratio of CPUs has decreased, making software-based RAID a viable alternative for system administrators in Linux-based environments.

backup copy of the partitioning scheme on each drive and installing GRUB (GRand Unified Bootloader) on the Master Boot Record (MBR) of each drive.

By keeping backup copies of drive partition tables, administrators can quickly restore an original partition table on a replacement drive and avoid having to edit configuration files or re-create partitions manually with the *fdisk* utility. To copy a partition table, administrators should create a directory in which to store the partition information, and then use the *sfdisk* command to write partition information files for each disk into that directory:

```
mkdir /raidinfo
sfdisk -d /dev/sda > /raidinfo/partitions.sda
(or hda for IDE drives)
sfdisk -d /dev/sdb > /raidinfo/partitions.sdb
(or hdb for IDE drives)
```

During the RAID configuration and OS installation process, the installer mechanism places GRUB on the MBR of the primary hard drive only (“*sda*” or “*hda*”). However, if the primary disk drive fails, the system can be booted only by using a boot disk. To avoid this problem, administrators should install GRUB on the MBR of each drive.

To enter the GRUB shell, type *grub* at the command prompt. Next, at the *grub>* prompt, type *find /grub/stage1*. The subsequent output will specify where the GRUB setup files are located. For example:

```
(hd0,0)
(hd1,0)
```

The output lists the locations of root for GRUB, which is GRUB syntax for where the */boot* partition is located. The Red Hat Linux OS specifically mounts the */boot* partition as the root partition for GRUB. In the example output, *sda* is *hd0* and *sdb* is *hd1* (these refer to SCSI drives; for IDE drives, *hda* is *hd0* and *hdb* is *hd1*). The second number specifies the partition number, where 0 is the first partition, 1 is the second partition, and so on. Thus, assuming SCSI disk drives, (*hd0,0*) signifies that the */boot* partition resides on the first partition of *sda* (*sda1*); (*hd1,0*) refers to */boot* residing on the first partition of *sdb* (*sdb1*).

Next, administrators should install GRUB on the MBR of the secondary RAID drive, so that if the primary drive fails, the next drive

has an MBR with GRUB ready to boot. When booting, the BIOS will scan the primary drive for an MBR and active partitions. If the BIOS finds them, it will boot to that drive; if not, it will go on to the secondary drive. Therefore, multiple drives in a system can have MBRs and active partitions, and the system will not have problems booting.

To install GRUB on the MBR of the secondary drive, administrators must temporarily define the secondary drive as the primary disk. To do so, administrators identify sdb (or hdb) as hd0, and instruct GRUB to write the MBR to it by typing the following at the prompt `grub>`:

```
device (hd0) /dev/sdb (or /dev/hdb for IDE drives)
root (hd0,0)
setup (hd0)
```

GRUB will echo all the commands it runs in the background of the `setup` command to the screen, and then will return a message that the `setup` command succeeded. Both drives now have an MBR, and the system can boot off either drive.

Identifying important Linux software RAID administration tools

After the system with software RAID is ready for production, several useful software RAID files and management utilities can help administrators manage the RAID devices:

- **/etc/raidtab:** This file contains information about the system's software RAID configuration, including which block devices belong to which md device. It can help administrators determine which RAID configuration the kernel expects to find on the system.
- **/proc/mdstat:** This file shows the real-time status of the md devices on the system, including online and offline partitions for each device. When rebuilding RAID partitions, this file also shows the status of that process.

In addition, the Red Hat Enterprise Linux `raidtools` package provides several useful tools:

- **lsraid:** This command-line tool allows administrators to list and query md devices in multiple ways. It presents much of the same information as `/etc/raidtab` and `/proc/mdstat`. Administrators can view this tool's man page for more information.

The 2.4 kernel of the Linux OS and its mature software RAID layer and management tools enable administrators to build an inexpensive RAID implementation.

- **raidhotadd:** This command-line utility allows administrators to add disk partitions to an md device and to rebuild the data on that partition.
- **raidhotremove:** This command-line utility allows administrators to remove disk partitions from an md device.

The next section demonstrates some key uses for these files and tools.

Restoring the RAID configuration after drive failure

When a drive in a RAID-1 array fails, administrators can restore the RAID array onto a new drive by following a three-step process: replace the failed drive, partition the replacement drive, and add the RAID partitions back into the md devices.

Replacing a disk drive

Once a hard disk drive fails, it must be replaced immediately to preserve the data redundancy that RAID-1 provides. The method by which the drive is replaced depends on the type of disk drives in the system. Because hot plugging of IDE drives is not supported in the Linux 2.4 kernel, administrators must replace an IDE drive by shutting down the system, swapping the drive, and then rebooting. However, the Linux device drivers for the Adaptec® and LSI Logic® SCSI controllers that ship on Dell™ PowerEdge™ servers do support hot plugging of drives, so administrators can replace SCSI drives while the system is still running.

To hot plug a SCSI disk drive, administrators first must disable the drive in the kernel, then physically replace the drive, and finally enable the new drive in the kernel. To disable a SCSI drive, administrators echo the device out of the real-time `/proc` file system within Linux, and the system instructs the corresponding drive to “spin down” and stop operating (for example, a 10,000 rpm drive would go from a speed of 10,000 rpm to 0 rpm). Conversely, to enable a SCSI drive, administrators echo the device into the real-time `/proc` file system, and the system instructs the drive to “spin up” to operating speed (using the previous example, the drive would go from 0 rpm to 10,000 rpm).

To obtain the syntax to pass to the `/proc` file system, administrators can type `cat /proc/scsi/scsi` at the command prompt. This command will provide a list of all SCSI devices detected by the kernel at the moment the command was received. For example, suppose a system has two SCSI disk drives, and the drive with SCSI ID 1 fails and must be replaced. The output of the `/proc/scsi/scsi` command would be:

```
Host:   scsi0 Channel: 00 Id: 00 Lun: 00
Vendor: Seagate Model: . . .
Host:   scsi0 Channel: 00 Id: 01 Lun: 00
Vendor: Seagate Model: . . .
```


To disable the drive in the kernel, the administrator would type the following command:

```
echo "scsi remove-single-device" 0 0 1 0 >
    /proc/scsi/scsi
```

The administrator then would receive a message stating that the system is spinning down the drive. Another message is sent when this process is complete, after which the administrator can remove the failed drive and replace it with a new one. To enable the new drive in the kernel, the administrator must spin it back up:

```
echo "scsi add-single-device" 0 0 1 0 >
    /proc/scsi/scsi
```

After this process completes, the kernel is ready to use the drive.

Partitioning the replacement drive

Once the failed disk drive has been replaced, administrators must restore the partitions that were saved earlier in the `/raidinfo` directory. For example, if replacing drive `sdb`, the administrator would issue the following command to restore the original partition scheme for `sdb` to the new drive:

```
sfdisk /dev/sdb < /raidinfo/partitions.sdb
```

Adding the RAID partitions back into the md device

Next, the administrator adds the partitions back into each RAID device. The `/proc/mdstat` file displays the status of each RAID device. For example, a system that is missing a partition from the `md0` device would show the following:

```
md0 : active raid1 sda1[0]
      40064 blocks [2/1] [U_]
```

This output indicates that `md0` is active as a RAID-1 device and that partition `sda1` is currently active in that RAID device. However, it also shows that the second partition is not available to the device, as denoted by the following information: the first line does not list a second partition; the output `[2/1]` denotes that two partitions should be available to the device (the first value), but only one is currently available (the second value); and the output `[U_]` shows that the second partition is offline.

To add partition `sdb1` back into the `md0` device and to rebuild the data on that partition, administrators use the following command:


```
raidhotadd /dev/md0 /dev/sdb1
```

While the partition is rebuilding, administrators can track the status by periodically viewing `/proc/mdstat`, which displays the percentage of rebuilding that is complete. Once the rebuilding is finished, `/proc/mdstat` would show the following output for the example device:

```
md0 : active raid1 sda1[0] sdb1[1]
      40064 blocks [2/2] [UU]
```

Administrators must complete the `raidhotadd` command to add each partition back into its respective RAID device. Once the failed drive has been replaced, administrators simply run the GRUB commands discussed in "Prepping the system for drive failure" to install GRUB on the MBR of the new disk. After this step, the RAID configuration will be fully restored. These functions can easily be placed into a script. Then, from a single executable point, administrators can complete all rebuild functions—making software RAID more palatable by easing drive administration.

Building cost-efficient RAID in Linux

The increasing cost-effectiveness of software RAID offers Linux system administrators an alternative to more expensive hardware-based RAID implementations, thanks to the performance and cost advantages of the Linux OS and rapid advancements in processor power. Using the management tools available in Red Hat Enterprise Linux 3, administrators can create RAID implementations that best suit their data center requirements. 

John Hull (john_hull@dell.com) is a software engineer at Dell and is currently the lead Linux engineer for Dell Precision™ workstations.

Steve Boley (steve_boyey@dell.com) is a Gold Server Support senior network engineer at Dell. He provides hardware and software support for U.S.-based customers, with an emphasis on Linux. Steve is a Microsoft® Certified Systems Engineer (MCSE) and a Red Hat Certified Engineer.

FOR MORE INFORMATION

Dell and Linux:
<http://www.dell.com/linux>

Nadon, Robert and Thomas Luo. "Implementing Software RAID on Dell PowerEdge Servers." *Dell Power Solutions*, August 2003.
http://www.us.dell.com/content/topics/global.aspx/power/en/ps3q03_nadon?c=us&cs=55561-en&gs=biz

Introduction to TCP Offload Engines

By implementing a TCP Offload Engine (TOE) in high-speed computing environments, administrators can help relieve network bottlenecks and improve application performance. This article introduces the concept of TOEs, explains how TOEs interact with the TCP/IP stack, and discusses potential performance advantages of implementing TOEs on specialized network adapters versus processing the standard TCP/IP protocol suite on the host CPU.

BY SANDHYA SENAPATHI AND RICH HERNANDEZ

As network interconnect speeds advance to Gigabit Ethernet¹ and 10 Gigabit Ethernet², host processors can become a bottleneck in high-speed computing—often requiring more CPU cycles to process the TCP/IP protocol stack than the business-critical applications they are running. As network speed increases, so does the performance degradation incurred by the corresponding increase in TCP/IP overhead. The performance degradation problem can be particularly severe in Internet SCSI (iSCSI)-based applications, which use IP to transfer storage block I/O data over the network.

By carrying SCSI commands over IP networks, iSCSI facilitates both intranet data transfers and long-distance storage management. To improve data-transfer performance over IP networks, the TCP Offload Engine (TOE) model can relieve the host CPU from the overhead of processing TCP/IP. TOEs allow the operating system (OS) to move all TCP/IP traffic to specialized hardware on the network adapter while leaving TCP/IP control decisions to the host server. By relieving the host processor bottleneck, TOEs

can help deliver the performance benefits administrators expect from iSCSI-based applications running across high-speed network links. By facilitating file I/O traffic, TOEs also can improve the performance of network attached storage (NAS). Moreover, TOEs are cost-effective because they can process the TCP/IP protocol stack on a high-speed network device that requires less processing power than a high-performance host CPU.

This article provides a high-level overview of the advantages and inherent drawbacks to TCP/IP, explaining existing mechanisms to overcome the limitations of this protocol suite. In addition, TOE-based implementations and potential performance benefits of using TOEs instead of standard TCP/IP are described.

TCP/IP helps ensure reliable, in-order data delivery

Currently the de facto standard for internetwork data transmission, the TCP/IP protocol suite is used to transmit information over local area networks (LANs), wide area networks (WANs), and the Internet. TCP/IP

¹ This term does not connote an actual operating speed of 1 Gbps. For high-speed transmission, connection to a Gigabit Ethernet server and network infrastructure is required.

² This term does not connote an actual operating speed of 10 Gbps. For high-speed transmission, connection to a 10 Gigabit Ethernet (10GbE) server and network infrastructure is required.

processes can be conceptualized as layers in a hierarchical stack; each layer builds upon the layer below it, providing additional functionality. The layers most relevant to TOEs are the IP layer and the TCP layer (see Figure 1).

The IP layer serves two purposes: the first is to transmit packets between LANs and WANs through the *routing* process; the second is to maintain a homogeneous interface to different physical networks. IP is a connectionless protocol, meaning that each transmitted packet is treated as a separate entity. In reality, each network packet belongs to a certain data stream, and each data stream belongs to a particular host application. The TCP layer associates each network packet with the appropriate data stream and, in turn, the upper layers associate each data stream with its designated host application.

Most Internet protocols, including FTP and HTTP, use TCP to transfer data. TCP is a connection-oriented protocol, meaning that two host systems must establish a session with each other before any data can be transferred between them. Whereas IP does not provide for error recovery—that is, IP has the potential to lose packets, duplicate packets, delay packets, or deliver packets out of sequence—TCP ensures that the host system receives all packets in order, without duplication. Because most Internet applications require reliable data that is delivered in order and in manageable quantities, TCP is a crucial element of the network protocol stack for WANs and LANs.

Reliability. TCP uses the *checksum* error-detection scheme, which computes the number of set bits on packet headers as well as packet data to ensure that packets have not been corrupted during transmission. A TCP pseudoheader is included in the checksum computation to verify the IP source and destination addresses.

In-order data delivery. Because packets that belong to a single TCP connection can arrive at the destination system via different routes, TCP incorporates a per-byte numbering mechanism. This scheme enables the TCP protocol to put packets that arrive at their destination out of sequence back into the order in which they were sent, before it delivers the packets to the host application.

Flow control. TCP monitors the number of bytes that the source system can transmit without overwhelming the destination system with data. As the source system sends packets, the receiving system returns acknowledgments. TCP incorporates a *sliding window* mechanism to control congestion on the receiving end. That is, as the sender transmits packets to the receiver, the size of the window

As network speed increases, so does the performance degradation incurred by the corresponding increase in TCP/IP overhead.

reduces; as the sender receives acknowledgments from the receiver, the size of the window increases.

Multiplexing. TCP accommodates the flow from multiple senders by allowing different data streams to intermingle during transmission and receiving. It identifies individual data streams with a number called the TCP port, which associates each stream with its designated host application on the receiving end.

Traditional methods to reduce TCP/IP overhead offer limited gains

After an application sends data across a network, several data-movement and protocol-processing steps occur. These and other TCP activities consume critical host resources:

- The application writes the transmit data to the TCP/IP sockets interface for transmission in payload sizes ranging from 4 KB to 64 KB.
- The OS segments the data into maximum transmission unit (MTU)-size packets, and then adds TCP/IP header information to each packet.
- The OS copies the data onto the network interface card (NIC) send queue.
- The NIC performs the direct memory access (DMA) transfer of each data packet from the TCP buffer space to the NIC, and interrupts CPU activities to indicate completion of the transfer.

The two most popular methods to reduce the substantial CPU overhead that TCP/IP processing incurs are TCP/IP checksum offload and large send offload.

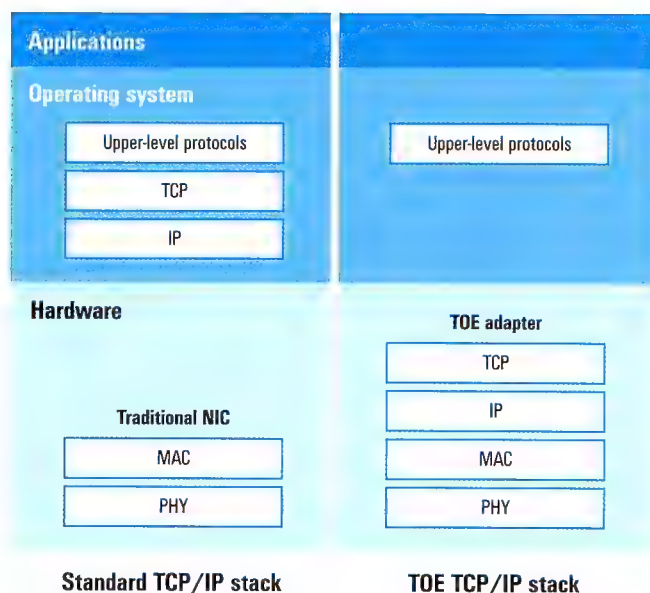


Figure 1. Comparing standard TCP/IP and TOE-enabled TCP/IP stacks

TCP/IP checksum offload

The TCP/IP checksum offload technique moves the calculation of the TCP and IP checksum packets from the host CPU to the network adapter. For the TCP checksum, the transport layer on the host calculates the TCP pseudoheader checksum and places this value in the checksum field, thus enabling the network adapter to calculate the correct TCP checksum without touching the IP header. However, this approach yields only a modest reduction in CPU utilization.

Large send offload

Large send offload (LSO), also known as TCP segmentation offload (TSO), frees the OS from the task of segmenting the application's transmit data into MTU-size chunks. Using LSO, TCP can transmit a chunk of data larger than the MTU size to the network adapter. The adapter driver then divides the data into MTU-size chunks and uses the prototype TCP and IP headers of the send buffer to create TCP/IP headers for each packet in preparation for transmission.

LSO is an extremely useful technology to scale performance across multiple Gigabit Ethernet links, although it does so under certain conditions. The LSO technique is most efficient when transferring large messages. Also, because LSO is a stateless offload, it yields performance benefits only for traffic being sent; it offers no improvements for traffic being received. Although LSO can reduce CPU utilization by approximately half, this benefit can be realized only if the receiver's TCP window size is set to 64 KB. LSO has little effect on interrupt processing because it is a transmit-only offload.

Methods such as TCP/IP checksum offload and LSO provide limited performance gains or are advantageous only under certain conditions. For example, LSO is less effective when transmitting several smaller-sized packages. Also, in environments where packets are frequently dropped and connections lost, connection setup and maintenance consume a significant proportion of the host's processing power. Methods like LSO would produce minimal performance improvements in such environments.

TOEs reduce TCP overhead on the host processor

In traditional TCP/IP implementations, every network transaction results in a series of host interrupts for various processes related to transmitting and receiving, such as send-packet segmentation and receive-packet processing. Alternatively, TOEs can delegate all processing related to sending and receiving packets to the network adapter—leaving the host server's CPU more available for business applications. Because TOEs involve the host processor only once for every application network I/O, they significantly reduce the number of requests and acknowledgments that the host stack must process.

Using traditional TCP/IP, the host server must process received packets and associate received packets with TCP connections, which means every received packet goes through multiple data copies from system buffers to user memory locations.

Because a TOE-enabled network adapter can perform all protocol processing, the adapter can use zero-copy algorithms to copy data directly from the NIC buffers into application memory locations, without intermediate copies to system buffers. In this way, TOEs greatly reduce the three main causes of TCP/IP overhead—CPU interrupt processing, memory copies, and protocol processing.

CPU interrupt processing

An application that generates a write to a remote host over a network produces a series of interrupts to segment the data into packets and process the incoming acknowledgments. Handling each interrupt creates a significant amount of context switching—a type of multitasking that directs the focus of the host CPU from one process to another—in this case, from the current application process to the OS kernel and back again. Although interrupt-processing aggregation techniques can help reduce the overhead, they do not reduce the event processing required to send packets. Additionally, every data transfer generates a series of data copies from the application data buffers to the system buffers, and from the system buffers to the network adapters.

High-speed networks such as Gigabit Ethernet compel host CPUs to keep up with a larger number of packets. For 1500-byte packets, the host OS stack would need to process more than 83,000 packets per second, or a packet every 12 microseconds. Smaller packets put an even greater burden on the host CPU. TOE processing can enable a dramatic reduction in network transaction load. Using TOEs, the host CPU can process an entire application I/O transaction with one interrupt. Therefore, applications working with data sizes that are multiples of network packet sizes will benefit the most from TOEs. CPU interrupt processing can be reduced from thousands of interrupts to one or two per I/O transaction.

Memory copies

Standard NICs require that data be copied from the application user space to the OS kernel. The NIC driver then can copy the data from the kernel to the on-board packet buffer. This requires multiple trips across the memory bus (see Figure 2): When packets are received from the network, the NIC copies the packets to the NIC buffers, which reside in host memory. Packets then are copied to the TCP buffer and, finally, to the application itself—a total of three memory copies.

A TOE-enabled NIC can reduce the number of buffer copies to two: The NIC copies

To improve data-transfer performance over IP networks, the TCP Offload Engine model can relieve the host CPU from the overhead of processing TCP/IP.

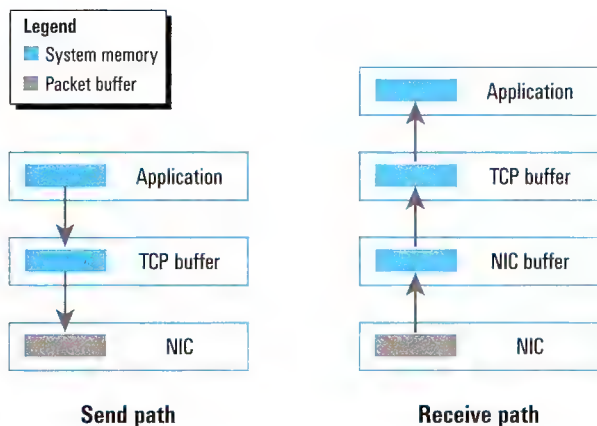


Figure 2. Transmitting data across the memory bus using a standard NIC

the packets to the TCP buffer and then to the application buffers. A TOE-enabled NIC using Remote Direct Memory Access (RDMA) can use zero-copy algorithms to place data directly into application buffers.

The capability of RDMA to place data directly eliminates intermediate memory buffering and copying, as well as the associated demands on the memory and processor resources of the host server—without requiring the addition of expensive buffer memory on the Ethernet adapter. RDMA also preserves memory-protection semantics. The RDMA over TCP/IP specification defines the interoperable protocols to support RDMA operations over standard TCP/IP networks.

Protocol processing

The traditional host OS-resident stack must handle a large number of requests to process an application's 64 KB data block send request. Acknowledgments for the transmitted data also must be received and processed by the host stack. In addition, TCP is required to maintain state information for every data connection created. This state information includes data such as the current size and position of the windows for both sender and receiver. Every time a packet is received or sent, the position and size of the window change and TCP must record these changes.

Protocol processing consumes more CPU power to receive packets than to send packets. A standard NIC must buffer received packets, and then notify the host system using interrupts. After a context switch to handle the interrupt, the host system processes the packet information so that the packets can be associated with an open TCP connection. Next, the TCP data must be correlated with the associated application and then the TCP data must be copied from system buffers into the application memory locations.

TCP uses the checksum information in every packet that the IP layer sends to determine whether the packet is error-free. TCP also records an acknowledgment for every received packet. Each of these operations results in an interrupt call to the underlying OS. As a

result, the host CPU can be saturated by frequent interrupts and protocol processing overhead. The faster the network, the more protocol processing the CPU has to perform.

In general, 1 MHz of CPU power is required to transmit 1 Mbps of data. To process 100 Mbps of data—the speed at which Fast Ethernet operates—100 MHz of CPU computing power is required, which today's CPUs can handle without difficulty. However, bottlenecks can begin to occur when administrators introduce Gigabit Ethernet and 10 Gigabit Ethernet. At these network speeds, with so much CPU power devoted to TCP processing, relatively few cycles are available for application processing. Multi-homed hosts with multiple Gigabit Ethernet NICs compound the problem. Throughput does not scale linearly when utilizing multiple NICs in the same server because only one host TCP/IP stack processes all the traffic. In comparison, TOEs distribute network transaction processing across multiple TOE-enabled network adapters.

TOEs provide options for optimal performance or flexibility

Administrators can implement TOEs in several ways, as best suits performance and flexibility requirements. Both processor-based and chip-based methods exist. The processor-based approach provides the flexibility to add new features and use widely available components, while chip-based techniques offer excellent performance at a low cost. In addition, some TOE implementations offload processing completely while others do so partially.

Processor-based versus chip-based implementations

The implementation of TOEs in a standardized manner requires two components: network adapters that can handle TCP/IP processing operations, and extensions to the TCP/IP software stack that offload specified operations to the network adapter. Together, these components let the OS move all TCP/IP traffic to specialized, TOE-enabled firmware—designed into a TCP/IP-capable NIC—while leaving TCP/IP control decisions with the host system. Processor-based methods also can use off-the-shelf network adapters that have a built-in processor and memory. However, processor-based methods are more expensive and still can create bottlenecks at 10 Gbps and beyond.

The second component of a standardized TOE implementation comprises TOE extensions to the TCP/IP stack, which are completely transparent to the higher-layer protocols and applications that run on top of them. Applications interact the same way with a TOE-enabled stack as they would with a standard TCP/IP stack. This transparency makes the TOE approach attractive because it requires no changes to the numerous applications and higher-level protocols that already use TCP/IP as a base for network transmission.

The chip-based implementation uses an application-specific integrated circuit (ASIC) that is designed into the network adapter. ASIC-based implementations can offer better performance than

off-the-shelf processor-based implementations because they are customized to perform the TCP offload. However, because ASICs are manufactured for a certain set of operations, adding new features may not always be possible. To offload specified operations to the network adapter, ASIC-based implementations require the same extensions to the TCP/IP software stack as processor-based implementations.

Partial versus full offloading

TOE implementations also can be differentiated by the amount of processing that is offloaded to the network adapter. In situations where TCP connections are stable and packet drops infrequent, the highest amount of TCP processing is spent in data transmission and reception. Offloading just the processing related to transmission and reception is referred to as partial offloading. A partial, or *data path*, TOE implementation eliminates the host CPU overhead created by transmission and reception.

However, the partial offloading method improves performance only in situations where TCP connections are created and held for a long time and errors and lost packets are infrequent. Partial offloading relies on the host stack to handle control—that is, connection setup—as well as exceptions. A partial TOE implementation does not handle the following:

- TCP connection setup
- Fragmented TCP segments
- Retransmission time-out
- Out-of-order segments

The host software uses dynamic and flexible algorithms to determine which connections to offload. This functionality requires an OS extension to enable hooks that bypass the normal stack and implement the offload heuristics. The system software has better information than the TOE-enabled NIC regarding the type of traffic it is handling, and thus makes offload decisions based on priority, protocol type, and level of activity. In addition, the host software is responsible for preventing denial of service (DoS) attacks. When administrators discover new attacks, they should upgrade the host software as required to handle the attack for both offloaded and non-offloaded connections.

TCP/IP fragmentation is a rare event on today's networks and should not occur if applications and network components are

By relieving the host
processor bottleneck,
TOEs can help deliver
the performance benefits
administrators expect
from iSCSI-based
applications running
across high-speed
network links.


working properly. A store-and-forward NIC saves all out-of-order packets in on-chip or external RAM so that it can reorder the packets before sending the data to the application. Therefore, a partial TOE implementation should not cause performance degradation because, given that current networks are reliable, TCP operates most of the time without experiencing exceptions.

The process of offloading all the components of the TCP stack is called full offloading. With a full offload, the system is relieved not only of TCP data processing, but also of connection-management tasks. Full offloading may prove more advantageous than partial offloading in TCP connections characterized by frequent errors and lost connections.

TOEs reduce end-to-end latency

A TOE-enabled TCP/IP stack and NIC can help relieve network bottlenecks and improve data-transfer performance by eliminating much of the host processing overhead that the standard TCP/IP stack incurs. By reducing the amount of time the host system spends processing network transactions, administrators can increase available bandwidth for business applications. For instance, in a scenario in which an application server is connected to a backup server, and TOE-enabled NICs are installed on both systems, the TOE approach can significantly reduce backup times.

Reduction in the time spent processing packet transmissions also reduces latency—the time taken by a TCP/IP packet to travel from the source system to the destination system. By improving end-to-end latency, TOEs can help speed response times for applications including digital media serving, NAS file serving, iSCSI, Web and e-mail serving, video streaming, medical imaging, LAN-based backup and restore processes, and high-performance computing clusters.

Part two of this article, which will appear in an upcoming issue, will include benchmark testing and analysis to measure the benefits of TOEs. 

Sandhya Senapathi (sandhya_senapathi@dell.com) is a systems engineer with the Server OS Engineering team. Her fields of interest include operating systems, networking, and computer architecture. She has an M.S. in Computer Science from The Ohio State University.

Rich Hernandez (rich_hernandez@dell.com) is a technologist with the Dell Product Group. He has been in the computer and data networking industry for more than 19 years. Rich has a B.S. in Electrical Engineering from the University of Houston and has pursued postgraduate studies at Colorado Technical University.

FOR MORE INFORMATION

RDMA Consortium:

<http://www.rdmaconsortium.org/home>

Improving Quality of Service

Using Dell PowerConnect 6024/6024F Switches

Quality of service (QoS) mechanisms classify and prioritize network traffic to improve throughput. This article explains the basic elements of QoS, focusing on how administrators can facilitate QoS with Dell™ PowerConnect™ 6024/6024F switches.

BY MARVELL SEMICONDUCTOR

Because network traffic can be so unpredictable, administrators often can provide only best-effort traffic delivery. Similar to the legal concept of a best-effort contract, best-effort delivery is considered to meet network service requirements—regardless of outcome—as long as the responsible party makes a genuine effort. However, network problems may cause loss, delay, or misdirection of traffic. Quality of service (QoS) mechanisms in network software help address this challenge by prioritizing packets to better meet required throughput levels.

Dell is helping to improve QoS on its computing platforms by adding flexibility and differentiation levels to its network switches. Dell™ OptiPlex™ and Dell Latitude™ PCs include integrated Gigabit Ethernet¹ connections that increase traffic flowing through the core switches. The Dell PowerConnect™ 6024 and 6024F switches can help provide advanced QoS and traffic engineering capabilities—including eight output queues, advanced queue mapping, and easy traffic metering—to meet the growing need for more advanced QoS services. This article explores hardware and software components of QoS and their implementation in PowerConnect 6024/6024F switches.

Given today's fast Ethernet switches operating with nonblocking multilayer switches, all at wire speed,

prioritizing traffic may no longer seem necessary. However, many IT organizations are reluctant to invest in expensive new Ethernet equipment. Converged voice and data service—which requires guaranteed minimum delays and assured bandwidth—increases network congestion even further. By implementing QoS, administrators can manage network congestion more effectively.

A QoS mechanism comprises three main elements that work together:

- **Access Control Lists (ACLs):** Both network security and QoS mechanisms apply ACL rules to determine what traffic can enter a switch, under what circumstances traffic can enter a switch, and what traffic is dropped. Only traffic that meets ACL criteria is subject to QoS settings.
- **Hardware queues:** The QoS mechanism assigns each packet to a hardware queue based on the traffic class to which the packet belongs.
- **Traffic-class handling attributes:** The QoS mechanism manages different traffic classes, such as bandwidth management, shaping, and policing, based on QoS attributes for each traffic class.

¹ This term does not connote an actual operating speed of 1 Gbps. For high-speed transmission, connection to a Gigabit Ethernet server and network infrastructure is required.

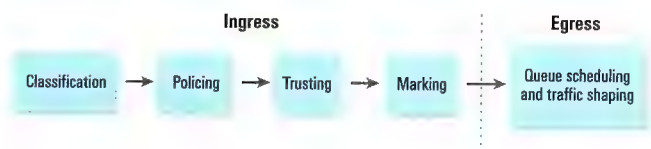


Figure 1. The QoS path for packets traversing the network switch

Ingress actions: Controlling packets

PowerConnect 6024/6024F switches provide three QoS modes, which govern the steps by which a packet traverses the switch (see Figure 1). Best-effort mode is equivalent to no QoS. Basic mode supports traffic classification on the ingress side and traffic shaping on the egress side. Advanced mode enables administrators to apply QoS actions such as policing, trusting, marking, and queue scheduling as well (see “Using advanced QoS mode to improve network traffic flow”).

Classifying packets determines service levels

Packets are first classified according to ACL rules specified by the network administrator. The classification process inspects packet fields, including IP address source and destination; IP subnet source and destination; TCP/UDP (Transmission Control Protocol/User Datagram Protocol) port information; virtual LAN (VLAN) ID; and Media Access Control (MAC) address source and destination. Packet classification helps determine what service level will be applied to a frame as it traverses the switch, and is identical to the classification performed for network security. Thus, the packet classification process results in either dropping the packets from the switch or forwarding them according to the specified ACL rules.

Policing packets controls bandwidth

Often, network traffic is asymmetrical. For example, Web users typically require ten times as much download bandwidth as upload bandwidth. Traffic policing (which is available only in advanced QoS mode) gives administrators the tools to control bandwidth precisely and deliver tiered services. Policing can be applied either to a specific traffic flow or to a group of traffic flows. In group assignment, or *aggregate policing*, administrators assign a traffic rate that governs all packets flowing in the designated group.

If a packet matches the predetermined profile, the policing function admits the packet to the network. If a packet does not fit the ACL profile, the policing function will either drop the packet immediately or admit the packet to the network according to the specified rules. If admitted, the packet is marked so that it will be handled

differently at downstream switches; its differentiated services code point (DSCP) is modified to lower the preferred status. The only limitations on policing are those imposed by the ACLs.

Alternatively, in basic mode, the DSCP of a packet can be rewritten in accordance with a DSCP-to-DSCP mutation map—a useful capability for a switch that resides at a QoS domain boundary. For example, one domain could use a given DSCP to designate a specific service level, while the domain to which this switch belongs uses different nomenclature. Rewriting the DSCP can retain service-level definitions between QoS domains.

The DSCP-to-DSCP translation applies only to ingress ports because they are DSCP-trusted. Applying this map to a port causes IP packets to be rewritten with newly mapped DSCP values at the ingress ports.

Policing uses a simple *token bucket algorithm*. This algorithm adds a token to the bucket as the switch receives each packet. The rate at which the bucket fills is a function of two factors: the rate at which tokens are removed, or *committed information rate* (CIR), and the bucket depth, or *committed burst size* (CBS). Packets that meet the CIR and CBS, called conforming packets, continue through the QoS path.

Trusting packets determines output service assignments

In advanced QoS mode, only conforming packets go through the trusting phase; nonconforming packets undergo an administrator-defined action. In basic mode, the administrator specifies “trust” for particular network domains. Within the trusted domain, the PowerConnect 6024/6024F marks each packet based on administrator-specified fields that signal the type of service the packet should receive. These fields also are used to assign the packet to one of eight output queues. The administrator determines the trust behavior by designating the fields upon which output service assignment is performed: 802.1p tag-based² fields, 802.1p port-based fields, Layer 3 predefined fields, and Layer 4 predefined fields.

Trusting in basic mode is a useful tool for switches at the edge of a QoS domain. In this case, the packets are classified at the edge of the domain and assigned to a queue. Sophisticated switches provide more flexible bandwidth management and control by supporting more queues. The PowerConnect 6024/6024F switch supports eight queues per port, allowing network administrators to configure eight differentiated levels of service for individuals and applications using the network.

In trusting (for either advanced or basic mode), the PowerConnect 6024/6024F switch assigns packets to a queue based on an administrator-configured map. The switch supports four different types of maps—class of service (CoS) to queue,

² The terms *class of service* and *802.1p tag* are used interchangeably.

USING ADVANCED QoS MODE TO IMPROVE NETWORK TRAFFIC FLOW

By offering three quality of service modes, Dell PowerConnect 6024/6024F switches help provide network administrators with the versatility to meet various networking needs. Although best-effort and basic QoS modes each offer advantages for managing network traffic, advanced QoS mode enables administrators to use policing, trusting, marking, and queue scheduling techniques to further streamline the flow of information through a network.

The following applications and network services provide examples of the traffic needs that administrators can address by implementing advanced QoS mode rather than simply increasing bandwidth:

- **H.323 voice over IP (VoIP) and telephony:** VoIP requires limited bandwidth but is sensitive to latency and jitter. VoIP applications include both call setup and control protocols; these traffic types are transmitted over TCP using known ports. Voice streaming is transmitted over UDP using a known port.
- **H.323 video:** Video is less delay-sensitive than VoIP but requires more bandwidth. Although the call setup and control protocols are the same as for voice and video streaming, H.323 video traffic uses a different UDP port.
- **Data transfer:** For network backup, data transfer requires significant bandwidth and low packet loss but has no specific delay requirements.

Figure A shows a typical network configuration to meet H.323 VoIP and telephony, H.323 video, and data transfer requirements.

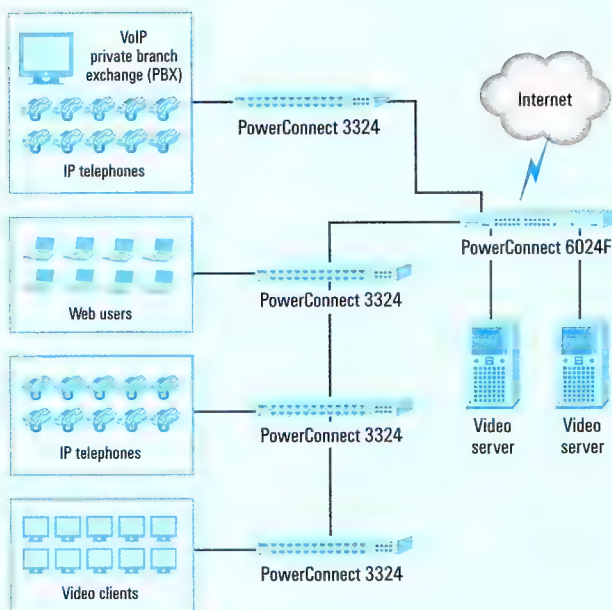


Figure A. Sample topology for a QoS environment

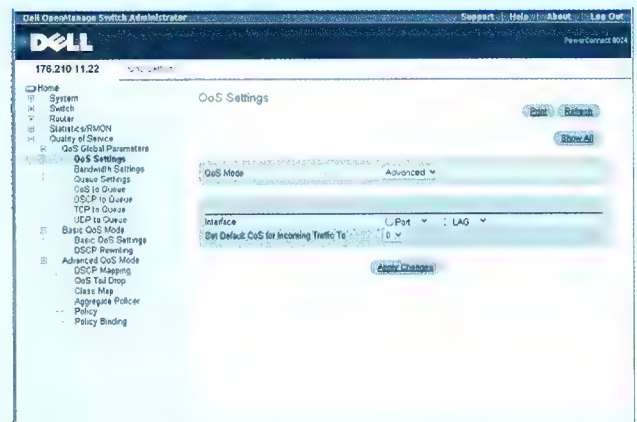


Figure B. QoS Settings screen: Advanced setting

The following example shows how administrators can ensure adequate QoS for a network with several applications, each presenting its own specific requirements. The example is based on the following assumptions:

1. The H.323 protocol uses TCP port 1720 for call setup (port number assigned by the Internet Assigned Numbers Authority, or IANA, for H.323) and port 1736 for audio call protocol.
2. UDP port 2776 is used for voice streaming.
3. UDP port 5004 is used for video streaming.
4. Input is derived from a Layer 3 switch that supports differentiated services code point (DSCP) tagging.
5. The voice data stream is tagged with the DSCP value of 63. (Its default mapping is to the highest priority queue—queue 8—like any value between 56 and 63.)
6. The video data stream DSCP and the Virtual Path Terminator (VPT) tags are unknown because neither is supported by the specific origination switch or subnet. For that reason, VPT or DSCP cannot be used as the trust parameter.

Configuring a PowerConnect 6024 switch

The following steps describe the process to configure a PowerConnect 6024 switch for the example scenario:

1. In the graphical user interface (GUI) provided with Dell OpenManage Switch Administrator, set the system to advanced mode on the QoS Settings screen shown in Figure B. Implement the following requirements using the advanced QoS mode of the PowerConnect 6024 switch:
 - a. Reserve enough bandwidth by limiting all other traffic to 700 Mbps—this is enough bandwidth for a worst-case scenario, protecting the network against faulty conditions.

- b. Prevent each real-time application from interfering with other real-time applications by limiting:

- Voice data stream to 60 Mbps
- Video data stream to 200 Mbps
- H.323 control to 5 Mbps

The committed bandwidth is adequate for the application. However, because these applications are higher priority, administrators must protect the lower-priority applications from unexpected traffic load of the higher-priority applications.

2. On the TCP to Queue screen, map the TCP call setup and control of H.323 to queues 5 and 6. Map TCP port 1736 to queue 6; after clicking the Apply Changes button, map TCP port 1720 to queue 5. Queues 5 and 6 are lower priority than queue 8 (used for voice) and queue 7 (used for video). The mapping of voice and video follows the default DSCP mapping table.
3. On the QoS Aggregate Policer screen, create a policy aggregator to limit all non-real-time traffic to 700 Mbps. In this example, the Policer is named "agany" and assigned an ingress committed information rate (CIR) of 700,000 Kbps and an ingress committed burst size (CBS) of 1,400,000 bytes.
4. On the Add ACE to IP-Based ACL screen, create five ACLs with the associated access control entries (ACEs). An ACL is created by using the Add button to add an ACL with the first associated ACE rule. A new ACE rule is added using the main screen. The five ACLs include:
 - a. Web traffic ACL: First ACE permits any traffic with IP destination port 80 (the general assigned value used for HTTP by the IANA); second ACE permits traffic to TCP port 20 (the general assigned value used for file transfer by the IANA); and third ACE permits traffic to TCP port 21
 - b. All non-real-time traffic ACL
 - c. Voice data stream ACL: Permits traffic with IP destination port 2776
 - d. Video data stream ACL: Permits traffic with IP destination port 5004

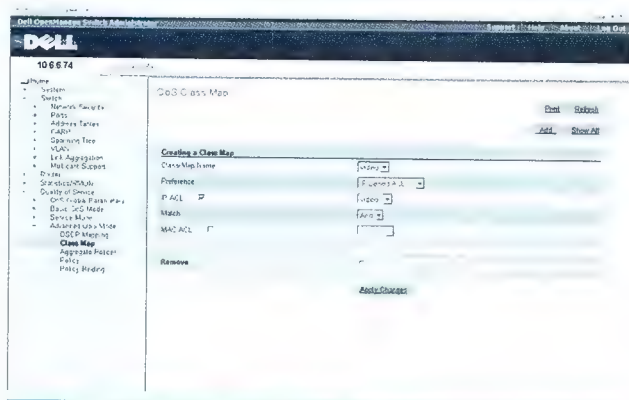


Figure C. QoS Class Map screen

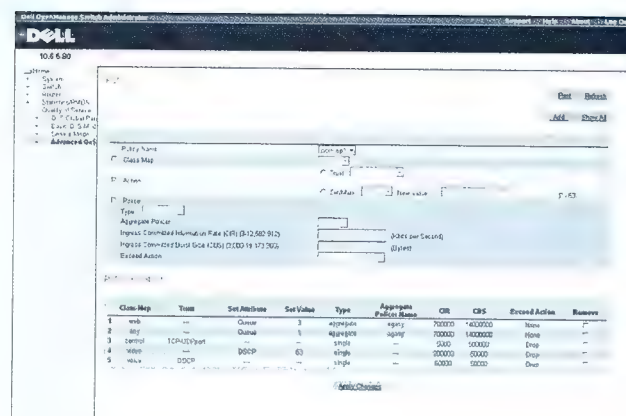


Figure D. Policy screen

- e. H.323 control ACL: First ACE permits traffic with IP destination port 1720; and second ACE permits traffic with IP destination port 1736
5. On the QoS Class Map screen shown in Figure C, bind the ACLs to five class maps. In advanced QoS mode, ACLs are not bound directly to interfaces and must be included in a policy, which includes a class map.
6. On the Policy screen shown in Figure D, create a new policy using the Add button. In this example scenario, the policy is named "polmap1" and includes the five class maps. Administrators define the class maps by completing the form on this screen for each field of information. Bind all the class maps to the policy and perform the following operations on the different classes:

Video class

- a. Limit the traffic using the Policer to 200 Mbps, reserving adequate bandwidth for the video traffic.
- b. Change the DSCP frames tag value classified by the video class to 53 (marking assumption 6). The default queue for this DSCP value is queue 7.

Voice class

- a. Limit the traffic using the Policer to 60 Mbps, reserving adequate bandwidth for the voice traffic.
- b. Use DSCP Trust mode to set DSCP queuing policy for all voice-classified frames. The default queue for DSCP 63 is queue 8.

Control class

- a. Limit the traffic using the Policer to 5 Mbps, reserving adequate bandwidth for the H.323 control.
- b. Use the TCP/UDP command to set the queuing policy for classified TCP/UDP frames.

Web class

- a. Direct all classified frames to queue 3 (lower-priority queue).
- b. Use the Policer to limit traffic according to the aggregated policy.

(continued on following page)

USING ADVANCED QoS MODE TO IMPROVE NETWORK TRAFFIC FLOW, *continued***Other traffic classes**

- a. Direct all classified frames to queue 1 (the lowest queue).
- b. Use the Policer to limit traffic according to the aggregated policy.

Figure D shows the full content of this policy after it has been completely defined.

7. On the QoS Policy Binding screen, attach the "polmap1" policy to the interface.

DSCP to queue, TCP to queue, and UDP to queue—allowing packets to be prioritized based on Layer 2, Layer 3, or Layer 4 behaviors.

Marking packets assigns administrator-defined actions

If a packet does not conform to policing rules, then it is marked for further action. DSCP values of nonconforming packets marked with the administrator-defined action *policed-dscp-transmit* will be rewritten according to the policed-DSCP map.

Egress actions: Controlling network congestion

Network congestion occurs when packets arrive at an output port faster than they can be transmitted. Switches use two general methods for controlling congestion in the network: queue scheduling and traffic shaping.

Queue scheduling prioritizes packets

Queue scheduling allows administrators to control service-class access to a limited network resource: link bandwidth. By managing the amount of bandwidth allocated to each service class on an output port, administrators can help control network congestion.

Strict priority queuing. This type of queuing is a simple method for supporting differentiated service classes. Strict priority queuing first classifies packets by the switch and then places them in different priority queues. Packets from the highest priority queues are scheduled first. Within each queue, packets are scheduled in first in, first out (FIFO) order.

Weighted round-robin (WRR) queuing. WRR queuing addresses the limitations of the strict priority model by ensuring that lower-priority queues are not denied access to output buffer space and bandwidth. In WRR queuing, packets are first categorized into different service classes—FTP, multimedia, and voice—and then assigned to a specific queue dedicated to that particular class. Each queue is serviced in round-robin order.

WRR queuing efficiently supports the differentiated service classes for a manageable number of highly aggregated traffic flows. Administrators can implement WRR queuing in hardware and apply it to high-speed interfaces in both the core and the edge of enterprise networks.

This example scenario demonstrates how network administrators can use the advanced QoS mode of the PowerConnect 6024/6024F switches to manage network traffic. With the easy-to-use, Web-based Dell OpenManage Switch Administrator tool, administrators can use PowerConnect switches to share bandwidth resources among various types of applications, assigning adequate resources to each one while helping ensure that lower-priority applications are not starved for bandwidth.


Traffic shaping regulates packet flow

Traffic shaping smoothes packet flow and regulates the rate and volume of traffic entering the network. Traffic-shaping tools set limits on the token generator, token bucket, and packet queue length. Traffic that adheres to the token bucket parameters can be transmitted on the link, but traffic that does not conform to the administrator-specified profile remains in the queue until it does conform.

For example, a network administrator can assign a per-queue shaper and a per-port shaper. If a given flow meets the shaping criteria for a specific queue and passes through it, the flow will then be subject to the aggregate shaper on the port. Thus, packets may pass the queue shaper, but be dropped on the port shaper.

Managing quality of service

A key feature of the QoS application in PowerConnect 6024/6024F switches is ease of management. Dell OpenManage™ Switch Administrator provides an intuitive graphical user interface that enables administrators to assign services and treatment to traffic flow. An industry-standard command-line interface also helps administrators manage the switch.

QoS mechanisms help administrators solve the problems inherent in unpredictable network traffic by classifying and prioritizing incoming and outgoing packets. Using hardware that facilitates QoS, such as Dell PowerConnect 6024/6024F switches, IT administrators can help maximize their QoS implementations, relieve network congestion, and better meet guaranteed minimum delays. Dell OpenManage Switch Administrator further simplifies configuring switches for QoS. 

Marvell Semiconductor (<http://www.marvell.com>) is a leading global semiconductor provider of complete broadband communications and storage solutions. The company's diverse portfolio includes switch controllers and processors, transceivers, communications controllers, wireless devices, and storage products that power the communications infrastructure.

FOR MORE INFORMATION

Dell PowerConnect switches: <http://www.dell.com/networking>

10 Gigabit Ethernet

Helps Relieve Network Bottlenecks for Bandwidth-Intensive Applications

Los Alamos National Laboratory researchers configured standards-based Dell™ servers with Intel® PRO/10GbE LR Server Adapters to test the actual network throughput of 10 Gigabit Ethernet (10GbE)—based local area networks (LANs), metropolitan area networks (MANs), and wide area networks (WANs). This article provides an overview of those tests, highlighting the key components that can help create cost-effective, high-speed network connections for a wide range of server-centric applications.

BY MATT W. BAKER AND WU-CHUN FENG

As bandwidth-intensive applications propagate, 10 Gigabit Ethernet (10GbE)¹ adapters can provide a way to increase network throughput using cost-effective, standards-based servers that are already in place. The ratified 10GbE standard is the same as previous Ethernet standards in almost every respect. Ten Gigabit Ethernet is still Ethernet, ensuring interoperability with all existing Ethernet technologies. Although standards are being developed for moving 10GbE to copper wire, the current 10GbE standard—like early forms of previous Ethernet technologies such as Gigabit Ethernet²—runs over various types of fiber-optic media only. The current standard enables full-duplex 10GbE transmissions over distances up to 300 meters (0.16 mile) using multimode, 850-nanometer fiber-optic cable (SR); up to 10 kilometers (6 miles) using single-mode,

1,310-nanometer optical fiber (LR); and up to 40 kilometers (25 miles) using 1,550-nanometer fiber (ER).³

The recently ratified fiber-optic extensions to the Ethernet standard provide 10GbE with the capability to interoperate with Synchronous Optical Network (SONET) installations, paving the way to extend 10GbE backbones into metropolitan area networks (MANs) and wide area networks (WANs). At the same time, the 10GbE standard offers the potential to increase the performance and productivity of local area networks (LANs), as shown in Figure 1.

However, with each evolutionary step, the question of actual performance versus theoretical performance arises. If historical Ethernet developments are any guide, products based on the 10GbE standard will likely reach

¹ This term does not connote an actual operating speed of 10 Gbps. For high-speed transmission, connection to a 10GbE server and network infrastructure is required.

² This term does not connote an actual operating speed of 1 Gbps. For high-speed transmission, connection to a Gigabit Ethernet server and network infrastructure is required.

³ For more information, see "10 Gigabit Ethernet Technology Overview," Intel Corporation, at http://www.intel.com/network/connectivity/resources/doc_library/white_papers/pro10gbe_lr_sa_wp.pdf.

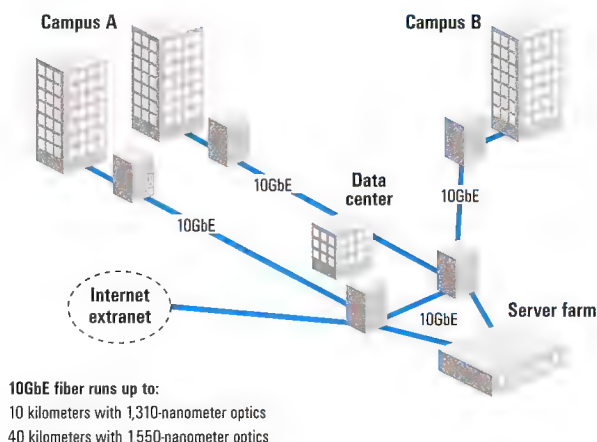


Figure 1. Configuring 10GbE links to expand LAN environments

their potential 10 Gbps transmission rate by 2005–2006. Meanwhile, IT managers pressed for more bandwidth must answer the immediate question: How fast is 10GbE right now?

Enhancing the productivity of server-based applications

Already, 10GbE network components are fast enough to satisfy many bandwidth-intensive applications. For example, to accelerate video rendering for *The Hulk*—a film released in 2003 that included a significant amount of computer-generated images—the filmmakers used a 10GbE trunk to create the conduit between servers in the artists' production network and servers in the production data center.

Meanwhile, researchers at Los Alamos and other national laboratories are using 10GbE components to power client/server data visualization applications. Los Alamos is using ParaView, an open source application designed to support distributed computing models that process large, complex data sets. Such applications help researchers analyze data sets in fields ranging from climatic modeling to DNA sequencing. Other applications enabled by 10GbE network components include *collaboratories*—which are based on a computing model that supports geographically dispersed collaborative research, particularly in scientific and engineering fields.⁴

Large-scale scientific applications benefit from 10GbE bandwidth because they typically require high-performance network connections for server-to-server as well as server-to-storage throughput. Using 10GbE bandwidth can help process distributed data up to six or seven times faster than Gigabit Ethernet in various applications. Another area in which 10GbE throughput can be beneficial is *bioinformatics*—collecting, classifying, storing, and analyzing biochemical and biological information. For example, in 2003 a

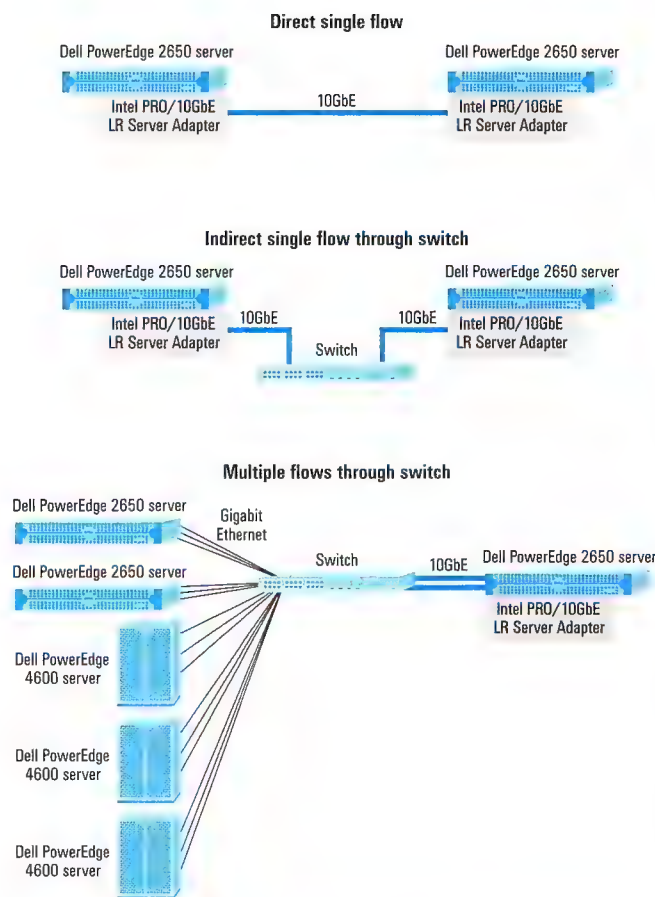


Figure 2. Configuring various server-to-server connections in a LAN

worldwide research effort took several weeks to identify the Severe Acute Respiratory Syndrome (SARS) virus. Had researchers been able to take advantage of 10GbE-linked collaboratories, they could have shared data worldwide with greater speed and efficiency, quite possibly resulting in faster identification of the virus.

Putting 10GbE to the test

To evaluate the actual performance of 10GbE network components today versus their theoretical performance potential, researchers at the Los Alamos National Laboratory tested a range of LAN, MAN, and WAN configurations.

LAN test configurations

The test team created several configurations for LAN testing, including direct single flow between two servers, indirect single flow between two servers through a switch, and multiple flows between servers through a switch (see Figure 2).⁵ Depending on the CPU, bus speed, and architecture of the servers used, Los Alamos

⁴ For more information about collaboratories, visit <http://www.accessgrid.org> or <http://www.scienceofcollaboratories.org>.

⁵ Standard Ethernet-compatible switches or routers can be used in any Ethernet configuration, including the configurations shown in Figure 2. However, these components must have 10GbE-compatible ports or adapters to make the necessary 10GbE server connections.

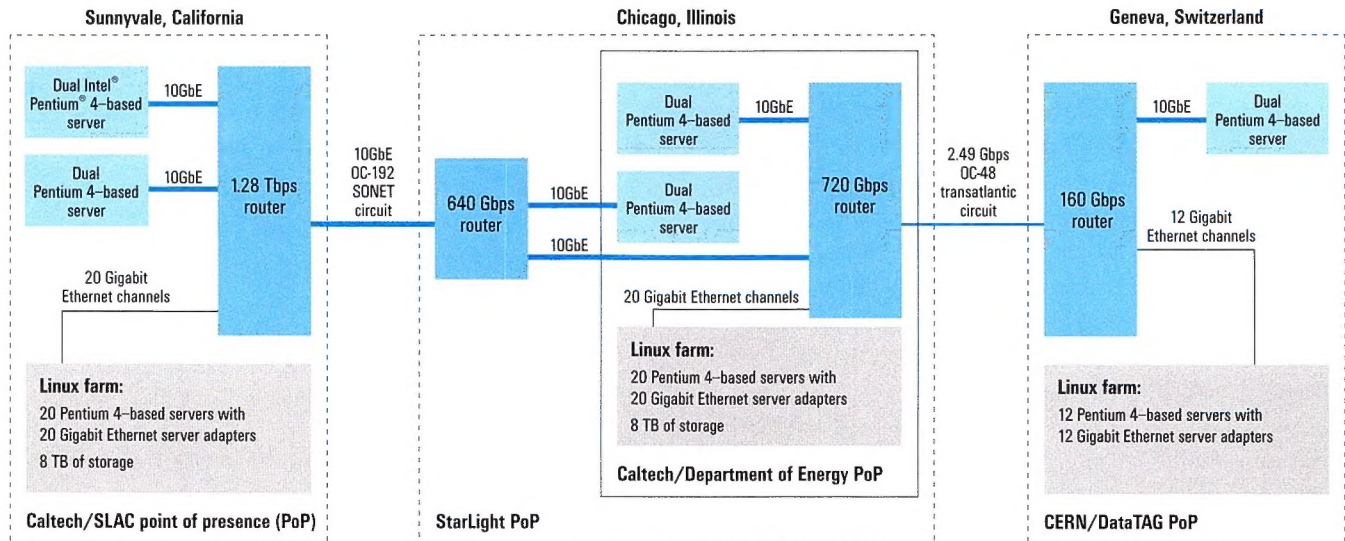


Figure 3. Configuring 10GbE WAN environment stretching from Sunnyvale, California, to Geneva, Switzerland

researchers were able to achieve an end-to-end throughput of over 7 Gbps between applications running on different Linux®-based servers, with end-to-end latency as low as 12 microseconds.⁶

MAN and WAN test configurations

For the MAN and WAN tests, Los Alamos researchers worked with the California Institute of Technology (Caltech), the Stanford Linear Accelerator Center (SLAC), and CERN (European Organization for Nuclear Research). This technical alliance tested 10GbE performance over the WAN configuration shown in Figure 3—a true wide area network that spanned more than 9,600 kilometers (6,000 miles) between Sunnyvale, California, and Geneva, Switzerland. Even though the 2.49 Gbps OC-48 transatlantic circuit presented a significant bottleneck, the WAN test group was able to transfer more than 1 TB of data in less than an hour with a sustained Sunnyvale-to-Geneva throughput of 2.38 Gbps⁶—breaking the then-current Internet2 Land Speed Record (I2-LSR) by 2.5 times in February 2003.⁷ More recently, in October 2003, a new I2-LSR record was set when Caltech and CERN researchers moved more than 1 TB of data across 7,000 kilometers (4,350 miles) in less than 30 minutes at a sustained throughput rate of 5.44 Gbps.⁸

Understanding key 10GbE network components

In both the Los Alamos LAN test configurations and the February 27, 2003, I2-LSR record-breaking configuration, the Intel® PRO/10GbE LR Server Adapter provided the Ethernet interface for Dell™ PowerEdge™ 2650 and PowerEdge 4600 servers. In the Los Alamos

configurations shown in Figure 2, the PRO/10GbE adapter helped to create high-performance TCP/IP-over-Ethernet network connections without requiring any modifications to the application code.

The PRO/10GbE adapter achieved its high performance through the architecture shown in Figure 4, which is based on the Intel 82597EX single-chip controller. This controller provides capabilities including direct memory access (DMA) without register mapping, minimized programmed I/O read access, and minimized interrupts for device management. In addition, the controller offloads various TCP/IP tasks such as checksums and segmentation from the host CPU.

The host server accesses the network adapter chip through the Peripheral Component Interconnect Extended (PCI-X®) interface, which connects to a 33/66 MHz, 32-/64-bit Peripheral Component Interconnect (PCI®) bus or a 33/66/100/133 MHz, 32-/64-bit PCI-X bus. To the right of the Media Access Control (MAC) layer is an 8B/10B physical encoding sublayer and a 10 Gbps media-independent interface (XGMII) for the 1,310-nanometer serial fiber-optics module.

The fiber-optics module provides optical transmission over single-mode fiber to distances of 10 kilometers. Although the PRO/10GbE adapter can support a 20 Gbps bidirectional data rate, current host PCI-X bus bandwidth presents at least one limiting factor: the peak network bandwidth of a 133 MHz, 64-bit PCI-X bus is 8.5 Gbps. Thus, with no other bottlenecks, 8.5 Gbps would be the maximum throughput to be expected from such servers. Fortunately, moving forward, new system interconnect technologies such as PCI Express™ will remove this artificial bottleneck.

⁶ "Optimizing 10 Gigabit Ethernet for Networks of Workstations, Clusters, and Grids: A Case Study" by Wu-chun Feng et al. in *Proceedings of ACM/IEEE SC 2003: High-Performance Networking and Computing Conference*, November 2003, <http://www.sc-conference.org/sc2003/paperpdfs/pap293.pdf>.

⁷ For more information see "I2-LSR Timeline, 27 February 2003" online at <http://lsr.internet2.edu/history.html>.

⁸ For more information, see "I2-LSR Timeline, 1 October 2003" online at <http://lsr.internet2.edu/history.html>.

First, the Los Alamos team tested unoptimized Dell PowerEdge 2650 servers, each with two 2.2 GHz processors, for single-flow TCP/IP throughput. The result using standard, 1500-byte maximum transfer units (MTUs) was 1.8 Gbps. When testers used a 9000-byte Jumbo frame MTU, they achieved network throughput of 2.7 Gbps. Although well short of 10 Gbps, 2.7 Gbps is a significant bandwidth advance over Gigabit Ethernet. Moreover, this throughput was accomplished using out-of-the-box PowerEdge 2650 servers.⁹

Fine-tuning network performance

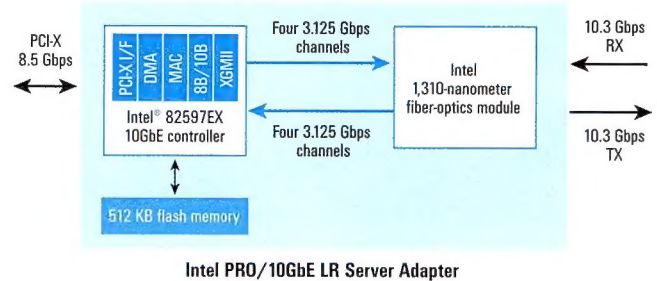
The Intel PRO/10GbE LR Server Adapter enables existing network servers to enter the 10GbE performance realm. Furthermore, it is highly customizable to meet particular application requirements for performance, network stability, and quality of service. To optimize existing servers, the Los Alamos team tested several approaches to isolate bottlenecks and increase bandwidth. These approaches included tuning TCP window sizes, increasing PCI-X burst size, and tuning MTU size.

For example, on PowerEdge 2650 servers using a standard TCP stack with 9000-byte MTUs, testers improved network throughput by 33 percent—from 2.7 Gbps to 3.6 Gbps—by increasing the PCI-X burst size to 4096 bytes. Even better performance was achieved by adjusting MTU sizes. Using an 8160-byte MTU, testers achieved a peak bandwidth of 4.11 Gbps.⁹ However, that result was accomplished running Linux and may be attributed to the Linux memory-allocation system; other operating systems may not achieve a comparable increase in throughput by using nonstandard MTUs.

To explore the network performance of high-end Linux-based servers, the Los Alamos test team ran preliminary trials on a uniprocessor server configured with an Intel® Itanium® 2 processor at 1.5 GHz and a PRO/10GbE LR Server Adapter. Using the same optimizations as with the PowerEdge 2650 system—an 8160-byte MTU and PCI-X burst size of 4096 bytes—the Itanium 2 processor-based server produced a unidirectional throughput of 7.2 Gbps with 12-microsecond socket-to-socket latency.^{9,10}

Extending the reach of networked applications

High bandwidth and low latency are the primary features of 10GbE network performance, helping to expand application capabilities, increase productivity, and reduce time to result for large-scale, complex scientific and engineering applications. In addition, enterprises that manipulate or store vast amounts of data may benefit from 10GbE performance in various applications, including feature-length film production, publishing, high-end graphics, finance, and economic modeling.¹⁰



Intel PRO/10GbE LR Server Adapter

Figure 4. Exploring the Intel PRO/10GbE LR Server Adapter architecture

IT administrators also can benefit from a 10GbE deployment—for example, faster network throughput enables faster backups. At full wire speed, a backup that takes one hour with Gigabit Ethernet can be completed in six minutes with 10GbE. But that is only the beginning. Emerging 10GbE technology promises the potential to provide a unified fabric interconnect for storage that could enable administrators to minimize or possibly eliminate costly proprietary network interconnects, which are traditionally used in such areas. Using standards-based Ethernet components as the network fabric could enable administrators to reduce equipment and maintenance costs significantly.

Moreover, 10GbE networks allow connections over significantly longer distances—as far as 10 kilometers of fiber-optic cable with 1,310-nanometer optics and 40 kilometers with 1,550-nanometer optics. These transmission distances can enable administrators to extend LANs across larger campuses and move LAN data centers to more cost-effective locations. ☺

Matt W. Baker (matt.w.baker@intel.com) is a technical marketing engineer for server network interface cards in the LAN Access Division at Intel.

Wu-chun Feng (feng@lanl.gov) leads the Research and Development in Advanced Network Technology (RADIANT) team at Los Alamos National Laboratory.

FOR MORE INFORMATION

Intel PRO/10GbE LR Server Adapter: http://www.intel.com/network/connectivity/products/pro10GbE_LR_server_adapter.htm

10 Gigabit Ethernet Technology Overview:
http://www.intel.com/network/connectivity/resources/doc_library/white_papers/pro10GbE_lr_sa_wp.pdf

Internet2 Land Speed Records:
<http://lsr.internet2.edu/history.html>

⁹ "Optimizing 10 Gigabit Ethernet for Networks of Workstations, Clusters, and Grids: A Case Study" by Wu-chun Feng et al. in *Proceedings of ACM/IEEE SC 2003: High-Performance Networking and Computing Conference*, November 2003, <http://www.sc-conference.org/sc2003/paperpdfs/pap293.pdf>.

¹⁰ For more information, see "Los Alamos National Lab Smashes Networking Records with Intel's 10 Gigabit Ethernet Server Adapter" online at http://www.intel.com/network/connectivity/case_studies/16832_LosAlamos_CS_r03.pdf.

Oracle Database 10g

\$149 Per User

One CD
17 minute install
Easy to use

Oracle Standard Edition One
\$149 per user or \$4995 per processor
First class database . . . economy price

ORACLE®

oracle.com/standardedition
or call 1.800.633.0753

Limitations and restrictions apply. Standard Edition One is available with Named User Plus licensing at \$149 per user with a minimum of five users or \$4995 per processor. Licensing of Oracle Standard Edition One is permitted only on servers that have a maximum capacity of 2 CPUs per server. 17 minute install is based upon testing on a system with 1x866MHz Intel CPU, 512 Mb RAM running Red Hat Linux 2.1. Actual install times will vary and are dependent on system configurations. For more information, visit oracle.com/standardedition

A photograph of two men in business suits sitting in the back of a car. They are both laughing heartily, looking towards the right. The man on the left is wearing a dark suit and a striped tie. The man on the right is wearing a light-colored suit and a patterned tie. The background shows a city street with buildings.

LINUX:

Operating system that
can now give you the same
buzz it gives IT.

There are tech reasons to love Linux. And there are business reasons. With new Novell® Nterprise™ Linux Services, you get both. An application suite that helps you boost productivity. And gives your tech staff the deep support they've been looking for. Bottom-line types will appreciate the increased efficiency that comes from running an array of applications – everything from secure identity management to remote access to messaging – on an operating system that's renowned for its low cost and reliability. And while everybody's in favor of better financial performance, what's really going to get your tech crew jazzed is the unprecedented multilevel support. Right on down to the actual operating system. To find out how new Novell Nterprise Linux Services can free your business to be more productive, call 1-800-218-1600 or visit www.novell.com/linux ☎ **WE SPEAK YOUR LANGUAGE.**

Novell.